Toward Rate-Distortion-Perception Optimality with Lattice Transform Coding

Shirin Saeedi Bidokhti
University of Pennsylvania
Joint work with Hamed Hassani and Eric Lei

Era of Massive High-Dimensional Data

Image/Video in autonomous systems

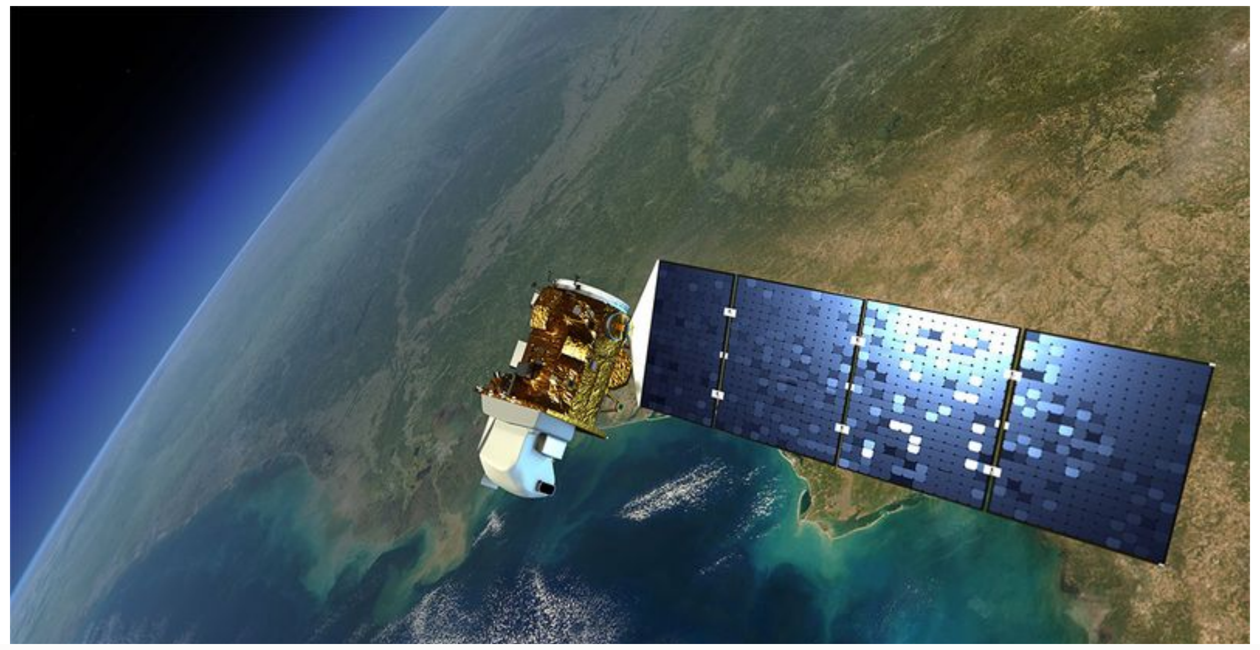
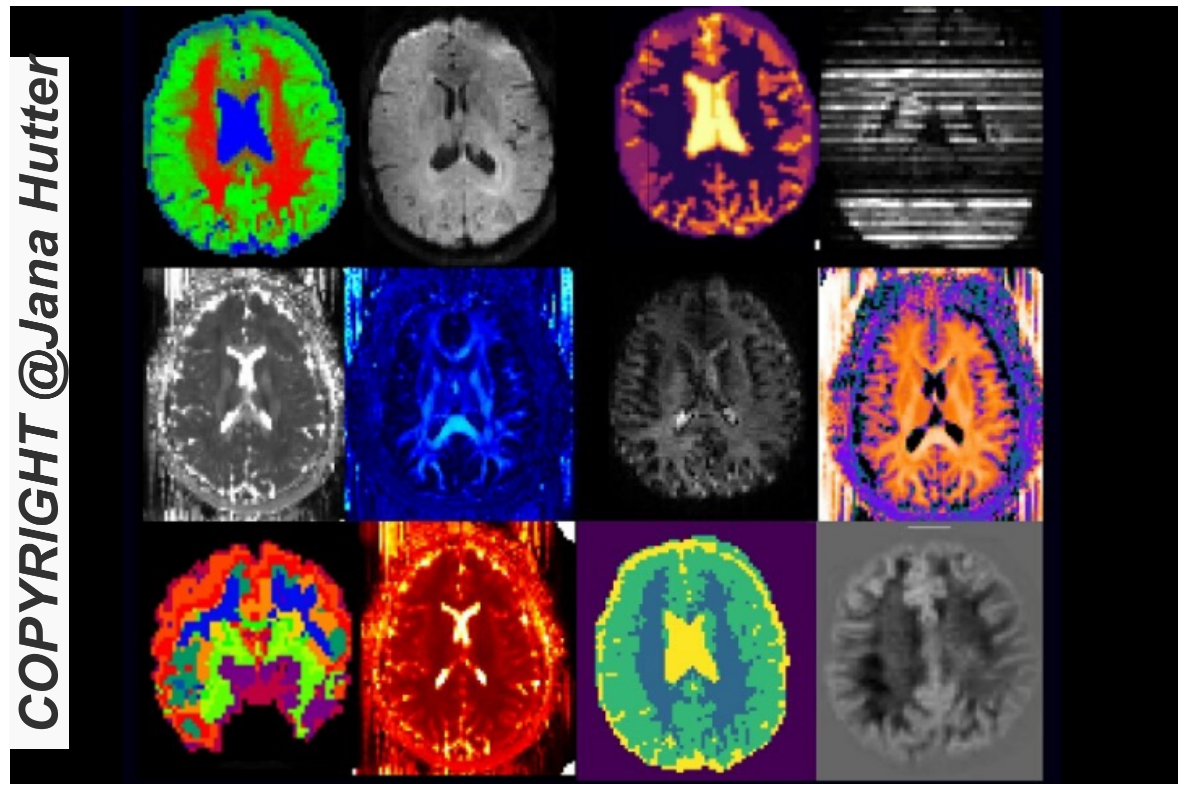


Image by NASA Goddard Space Flight Center via Flickr

Satellite and Remote Sensing Imagery

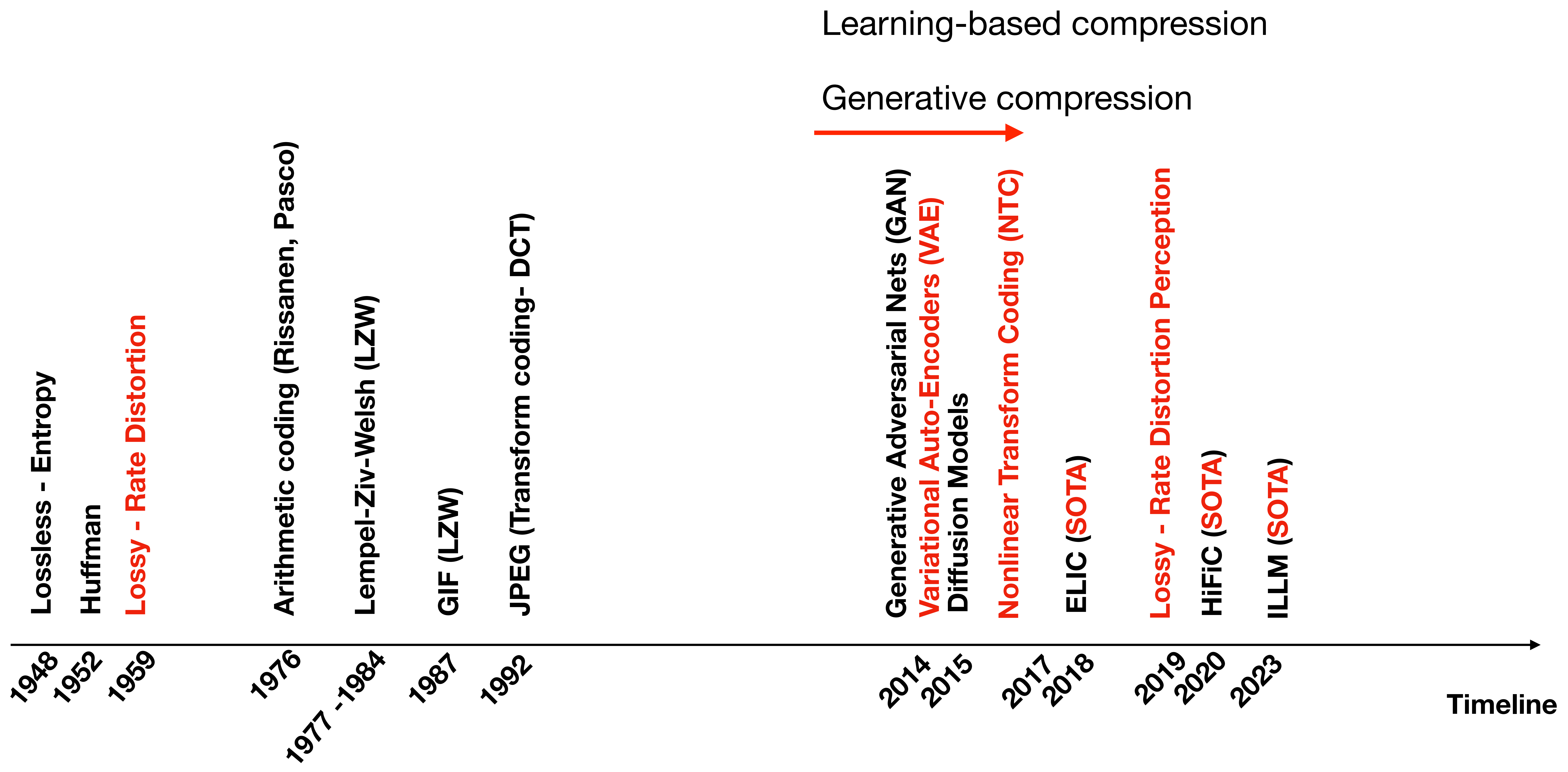


Medical imaging

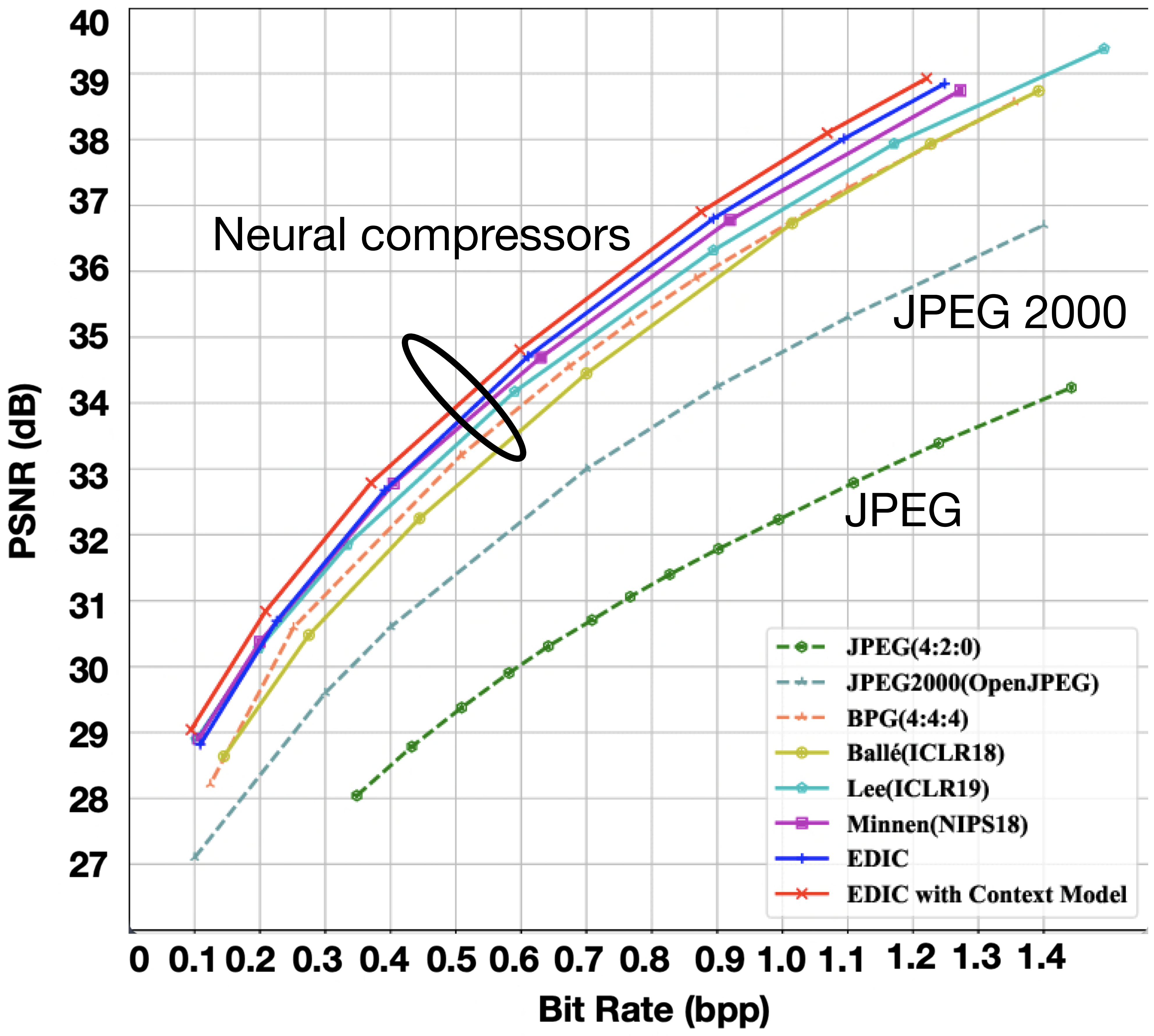
Graphical Scientific Datasets

- Data compression is critical for data storage, sharing, analysis

Data Compression Timeline



Success of Neural Compression



Proposed method, 3986 bytes (0.113 bit/px), PSNR: luma 27.01 dB/chroma 34.16 dB, MS-SSIM: 0.9039

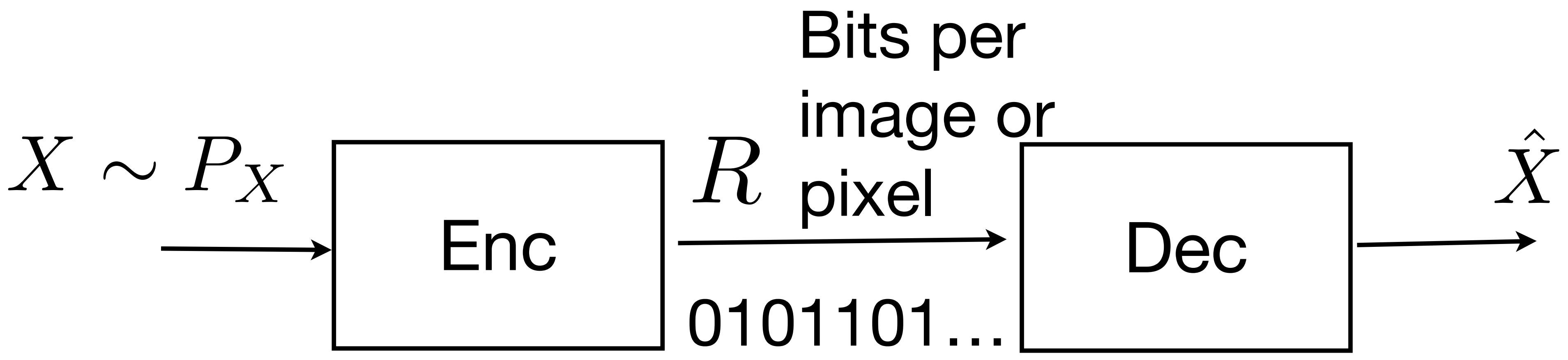
[Balle et al 2017]

JPEG, 4283 bytes (0.121 bit/px), PSNR: luma 24.85 dB/chroma 29.23 dB, MS-SSIM: 0.8079

JPEG

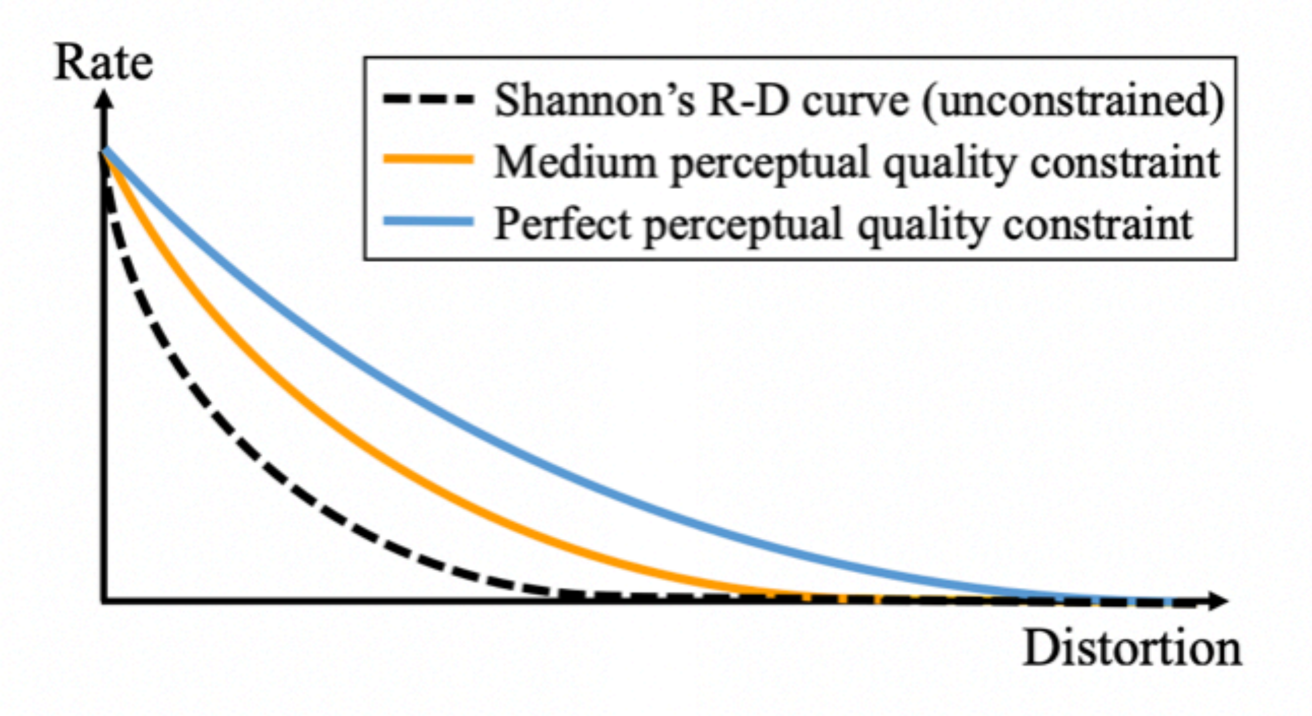
- Improved PSNR (distortion) for a given rate
- Improved perceptual quality
- It has further motivated the new theory of rate, distortion, perception

Rate-Distortion-Perception Function



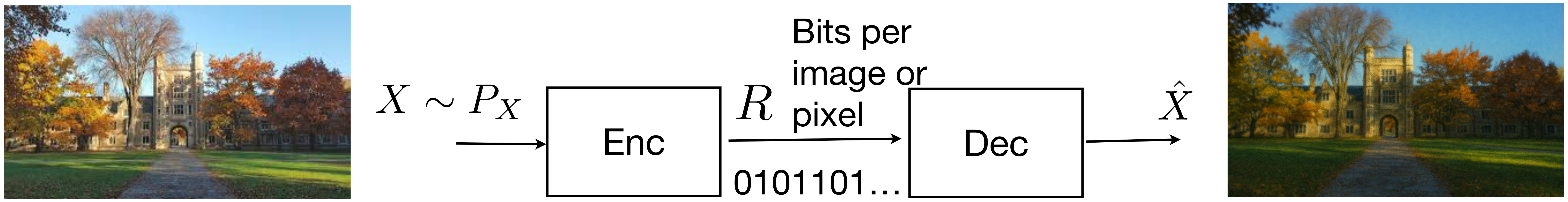
- Triple tradeoff between rate, distortion, perception [Blau&Michaeli ’19], [Matsumoto ’18], [Saldi et al ’15]
- RDP function:

$$R(D,P) = \min_{\substack{Q_{\hat{X}|X} \\ \mathbb{E}[d(X,\hat{X})] \le D \\ \delta(P_X, P_{\hat{X}}) \le P}} I(X;\hat{X})$$



Gaussian

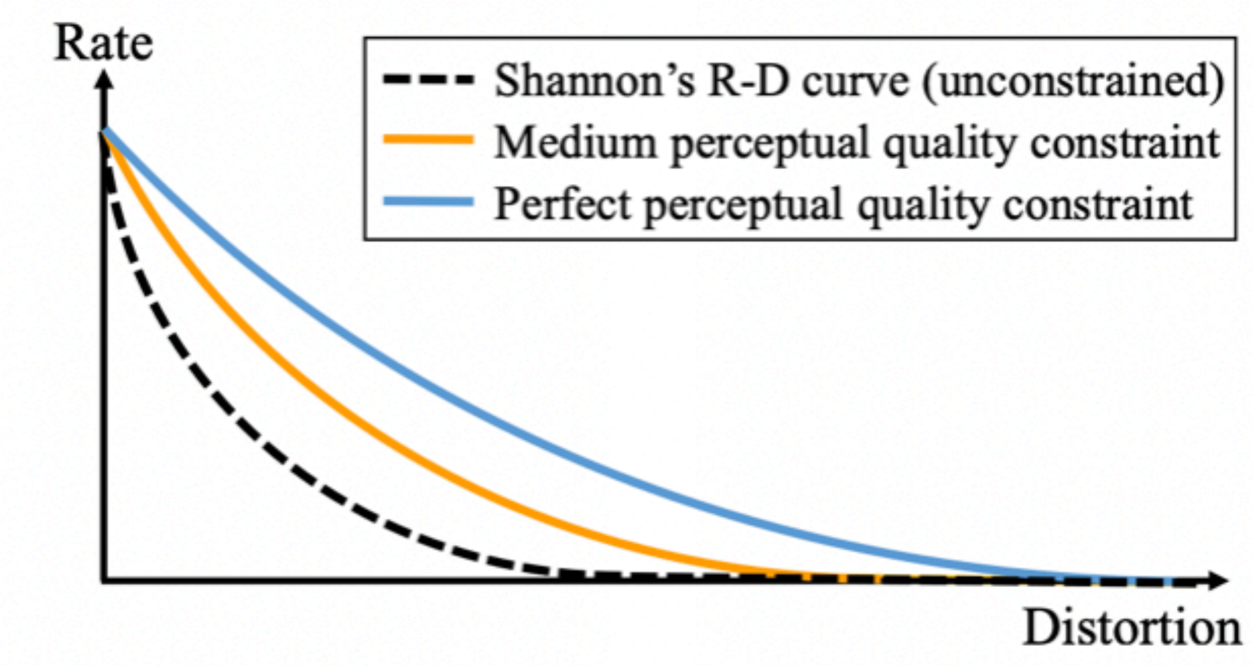
Rate-Distortion-Perception Function



- Triple tradeoff between rate, distortion, perception [Blau&Michaeli ’19], [Matsumoto ’18], [Saldi et al ’15]

- RDP function:

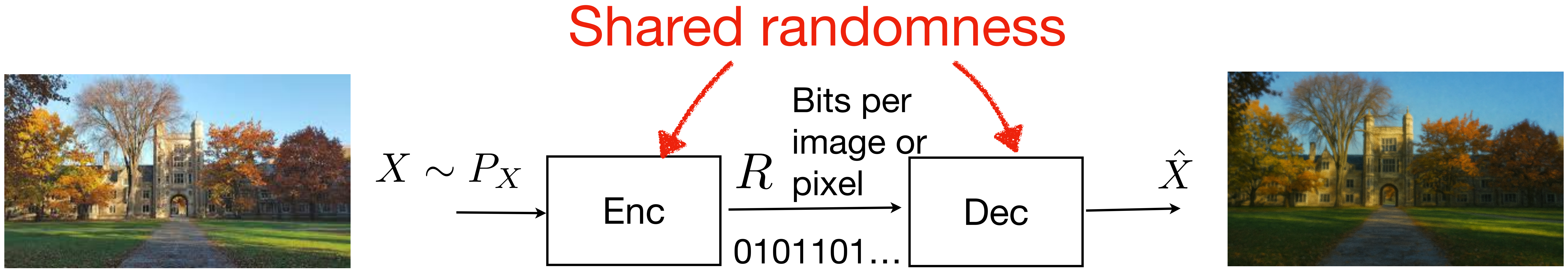
$$R(D,P) = \min_{\substack{Q_{\hat{X}|X} \\ \mathbb{E}[d(X,\hat{X})] \leq D \\ \delta(P_X, P_{\hat{X}}) \leq P}} I(X;\hat{X})$$



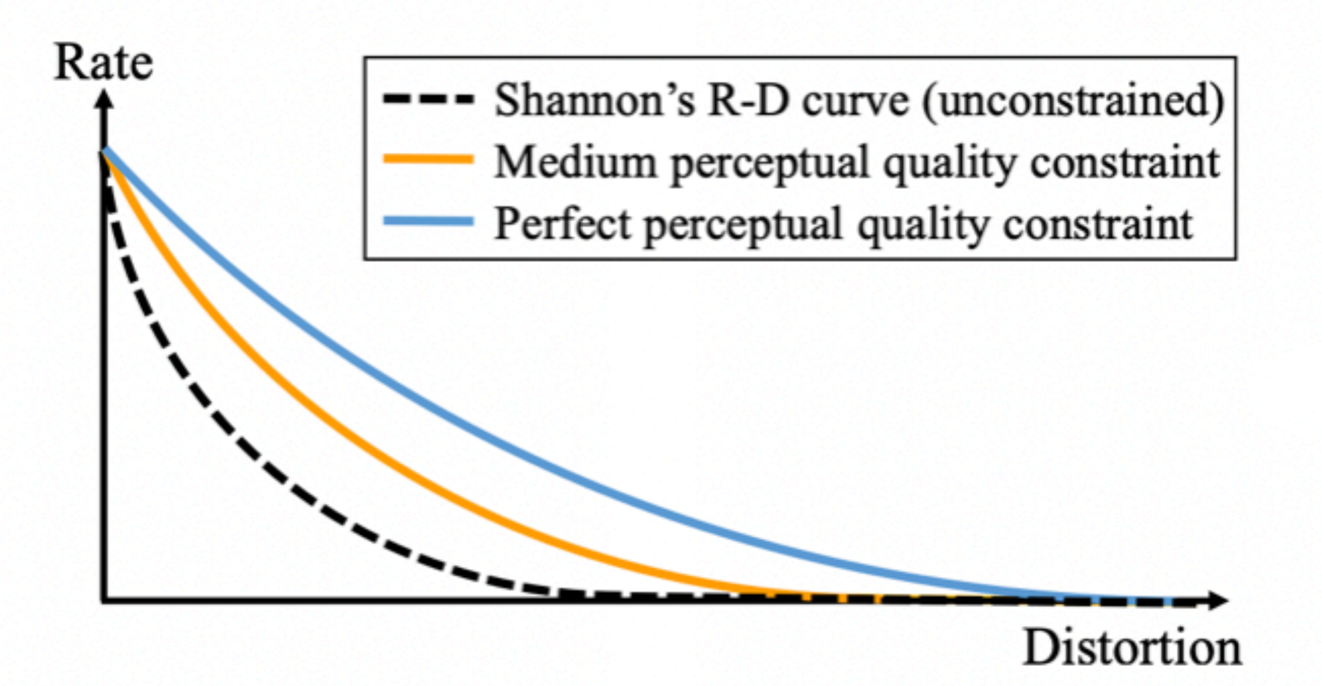
- RDP characterizes the fundamental limits of lossy compression under distortion and perception constrains [Theis&Wagner ’21]

Gaussian

Rate-Distortion-Perception Function

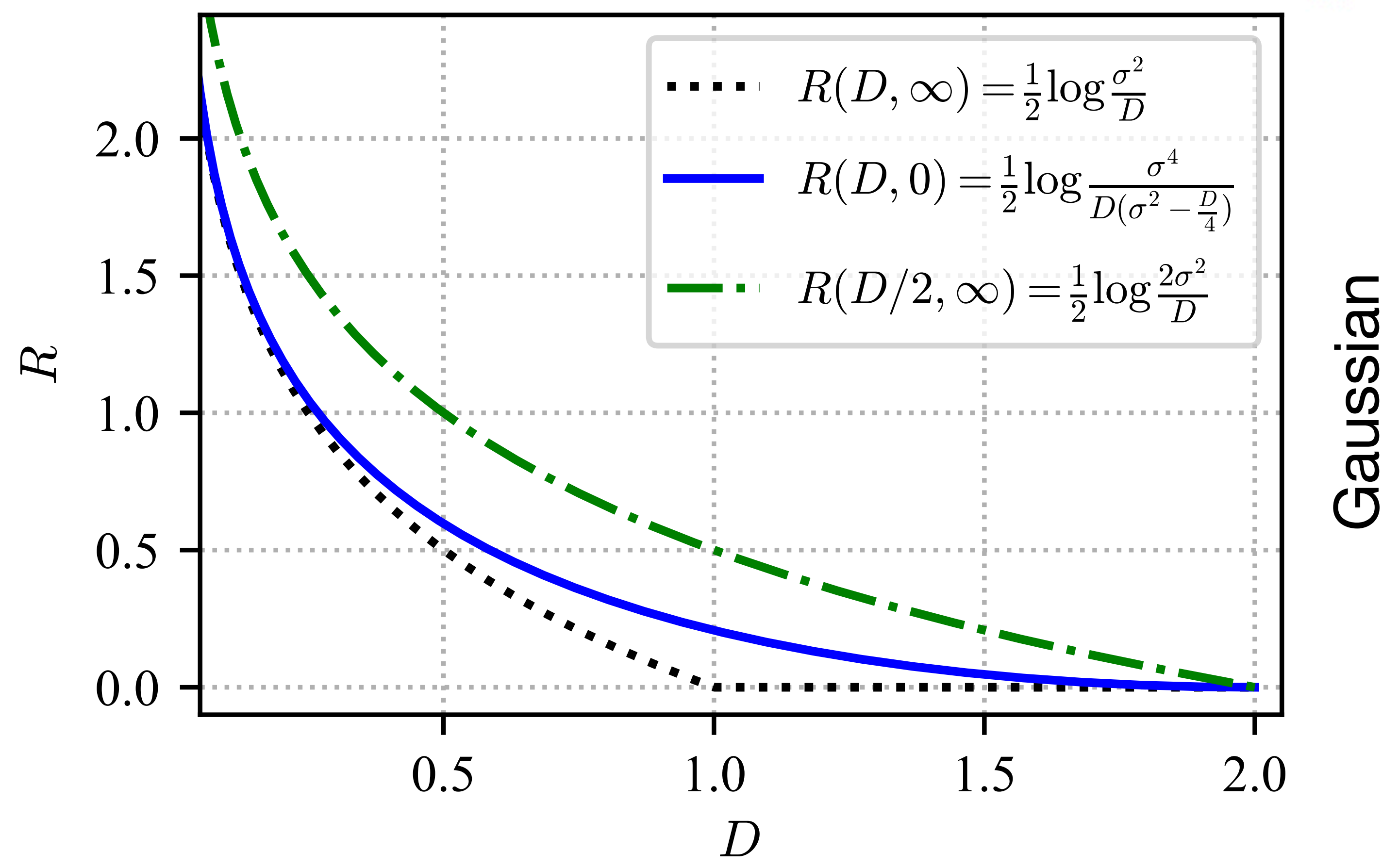


- Triple tradeoff between rate, distortion, perception [Blau&Michaeli '19], [Matsumoto '18], [Saldi et al '15]



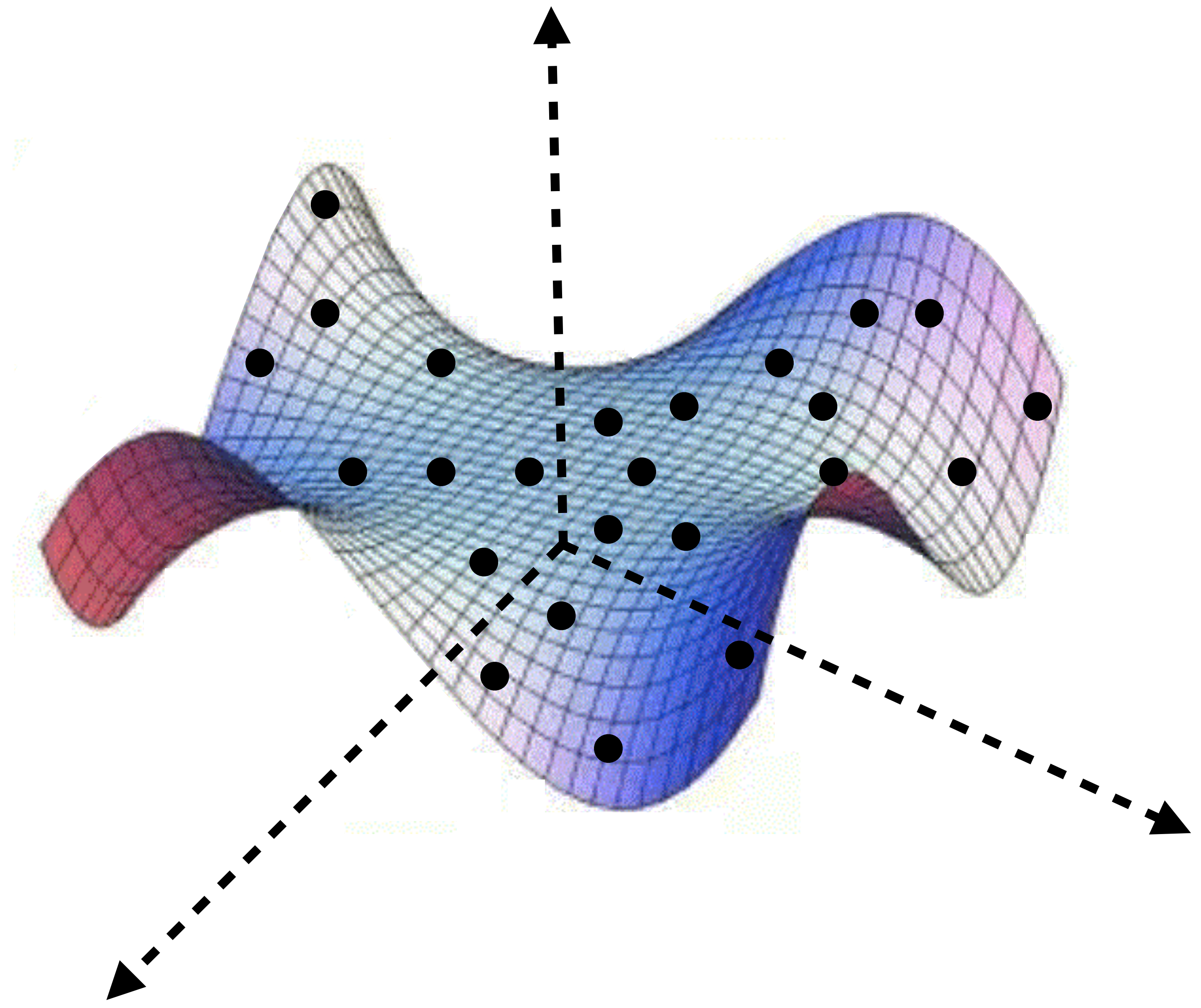
- RDP function: $$R(D,P) = \min_{\substack{Q_{\hat{X}|X} \\ \mathbb{E}[d(X,\hat{X})] \le D \\ \delta(P_X, P_{\hat{X}}) \le P}} I(X;\hat{X})$$

- RDP characterizes the fundamental limits of lossy compression under distortion and perception constrains [Theis&Wagner '21]
- Infinite shared randomness may be necessary [Saldi et al '15], [Chen et al '22], [Wagner '22]



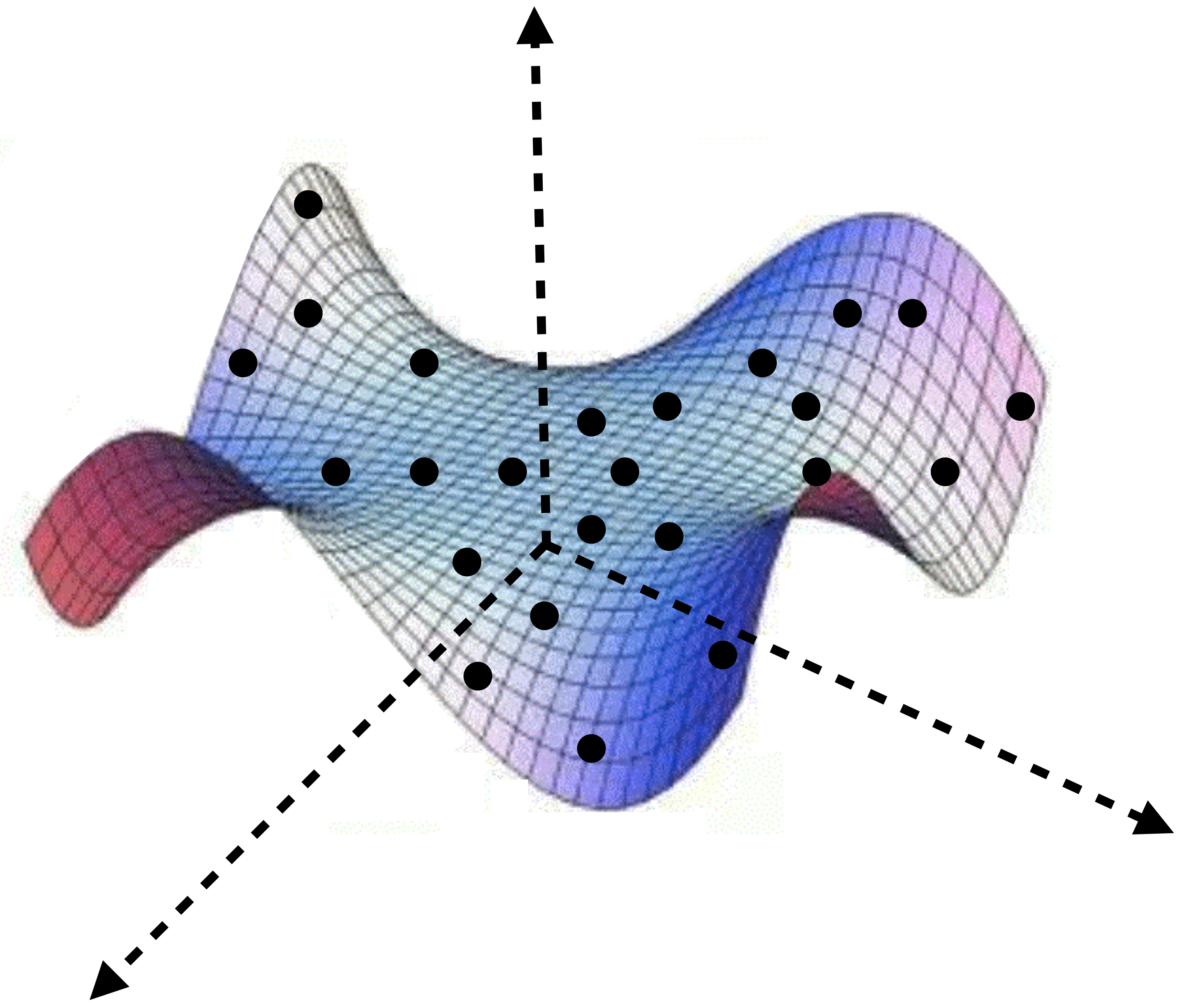
How is Learning Useful?

- Optimal schemes from information theory have exponential complexity in dimension
- Data is nominally high dimensional, but intrinsically is of much lower dimension



How is Learning Useful?

- Optimal schemes from information theory have exponential complexity in dimension
- Data is nominally high dimensional, but intrinsically is of much lower dimension
- the geometry:



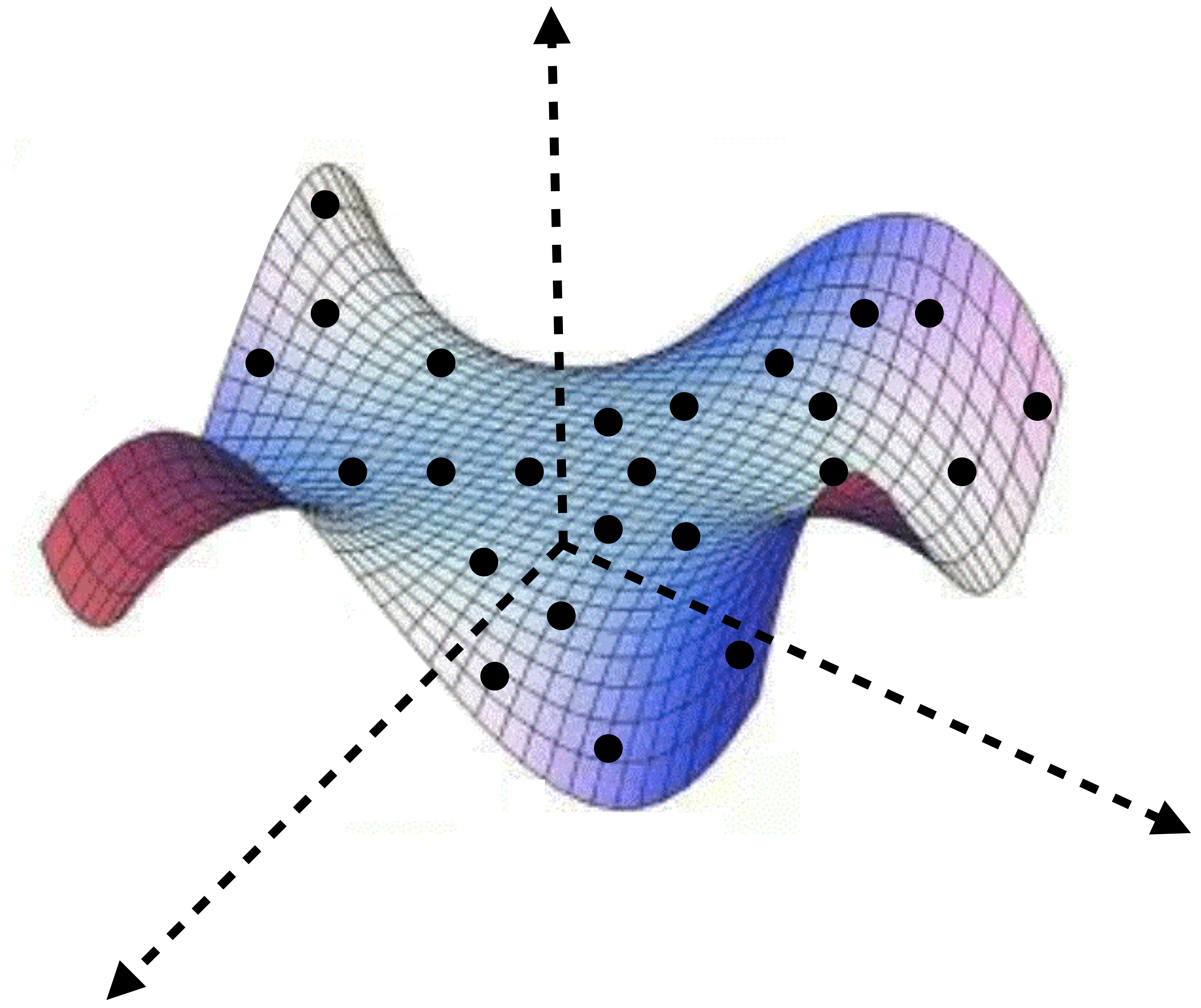
How is Learning Useful?

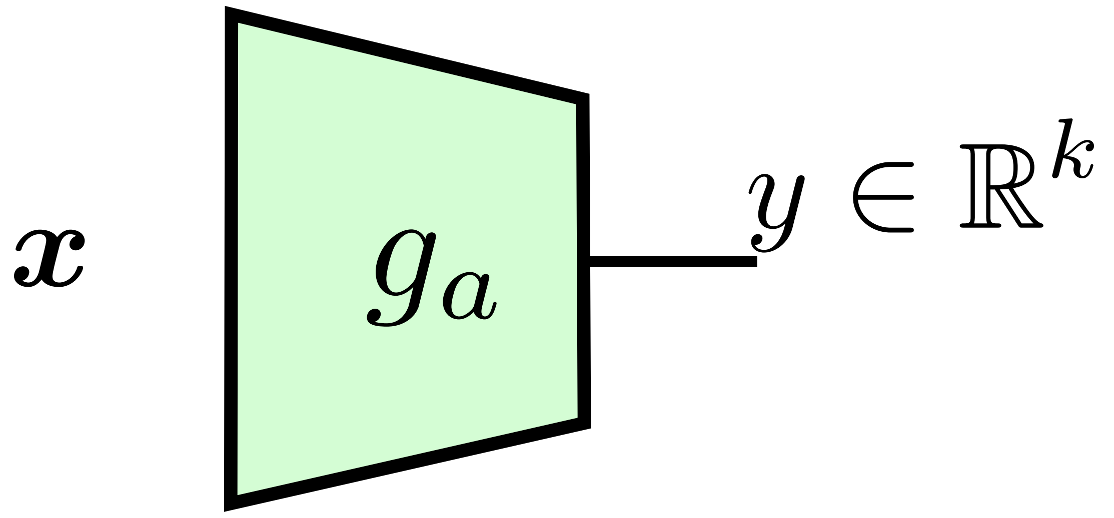
- Optimal schemes from information theory have exponential complexity in dimension

- Data is nominally high dimensional, but intrinsically is of much lower dimension

high-dim low-dim

- the geometry: $g_a : \mathbb{R}^n \to \mathbb{R}^k$

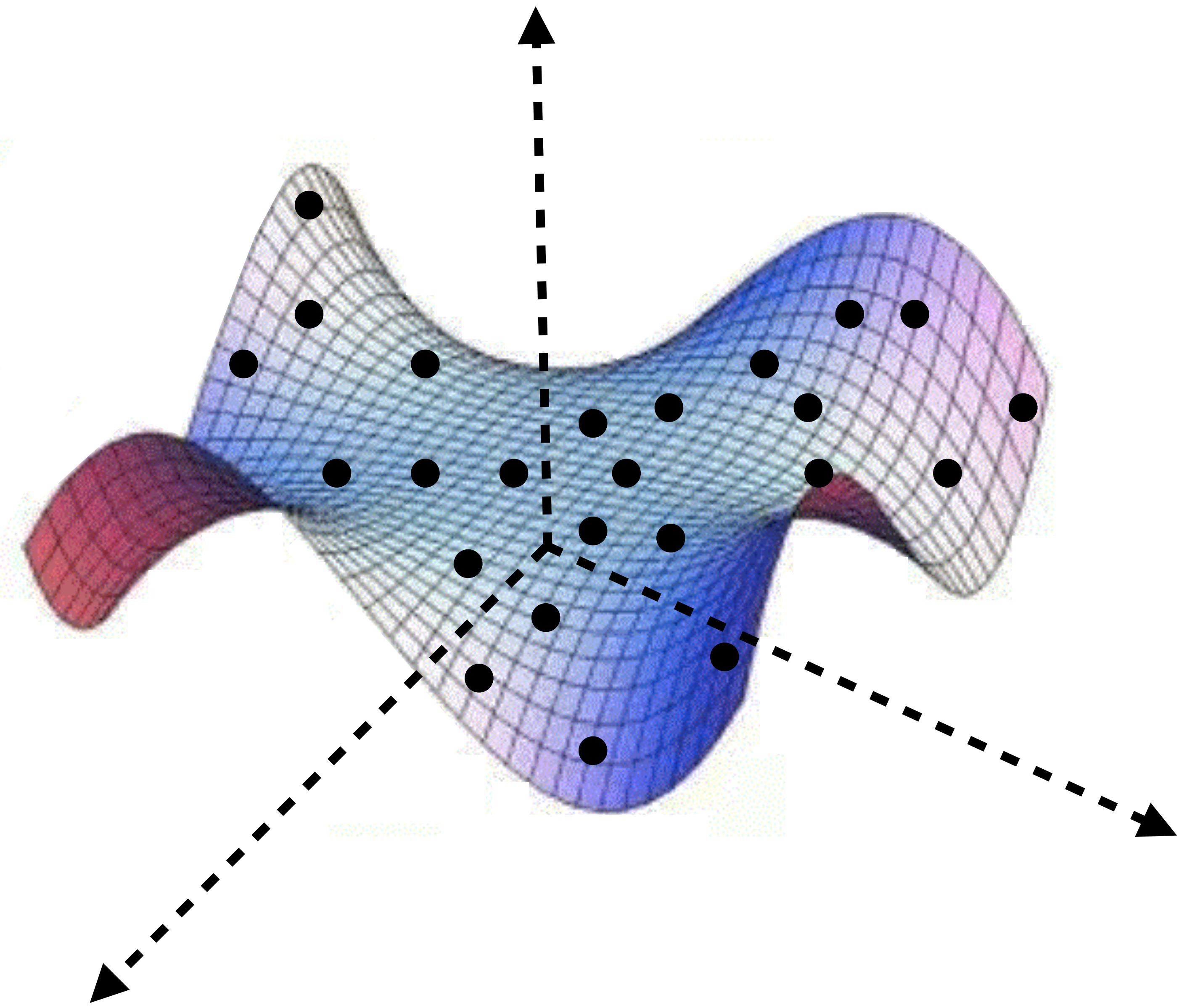


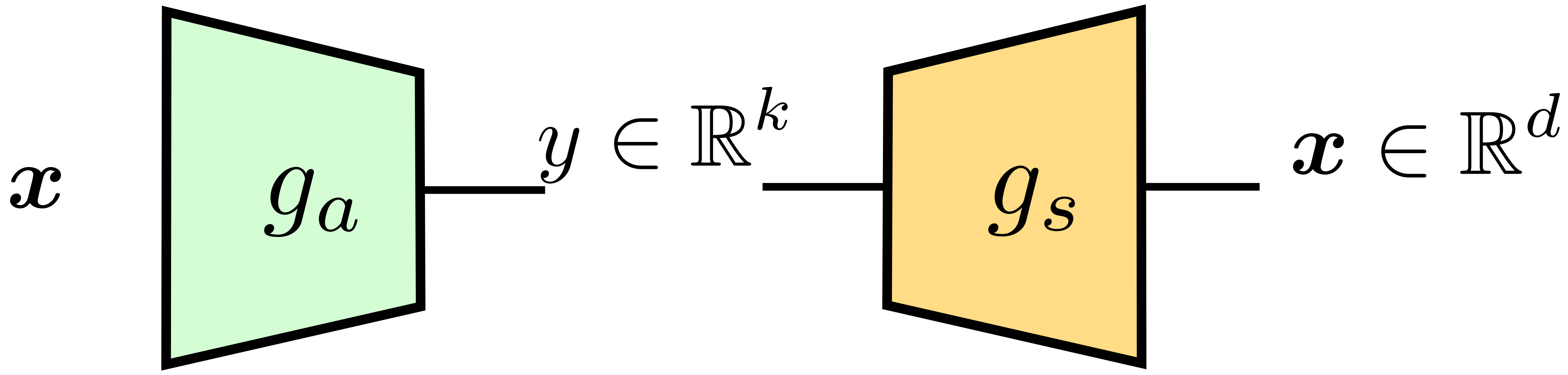


How is Learning Useful?

- Optimal schemes from information theory have exponential complexity in dimension
- Data is nominally high dimensional, but intrinsically is of much lower dimension
- the geometry: $g_a : \mathbb{R}^n \to \mathbb{R}^k$ (high-dim → low-dim)

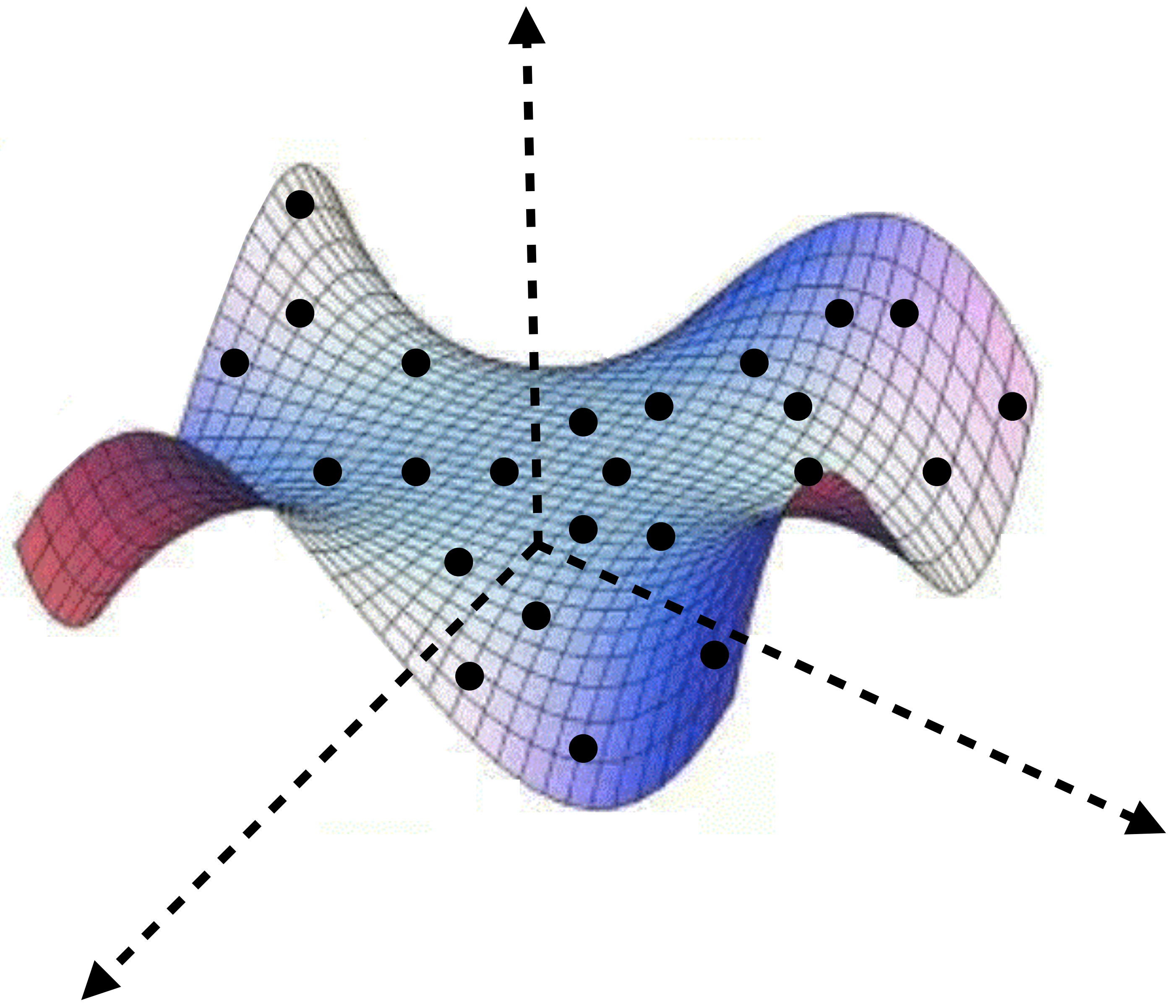
  $g_s : \mathbb{R}^k \to \mathbb{R}^n$ (low-dim → high-dim)





How is Learning Useful?

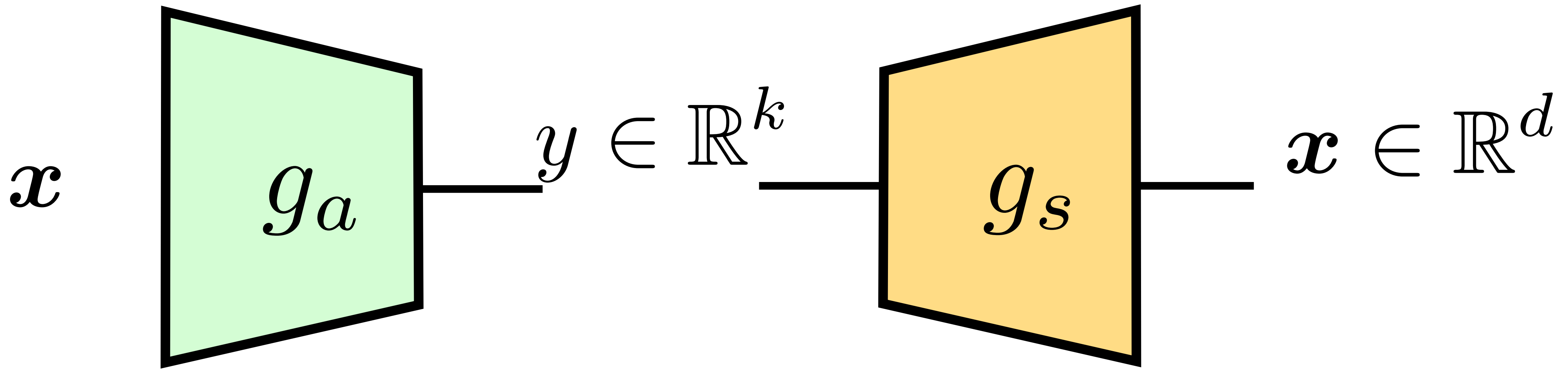
- Optimal schemes from information theory have exponential complexity in dimension
- Data is nominally high dimensional, but intrinsically is of much lower dimension



- the geometry: $\underset{\text{high-dim}}{g_a : \mathbb{R}^n} \to \underset{\text{low-dim}}{\mathbb{R}^k}$

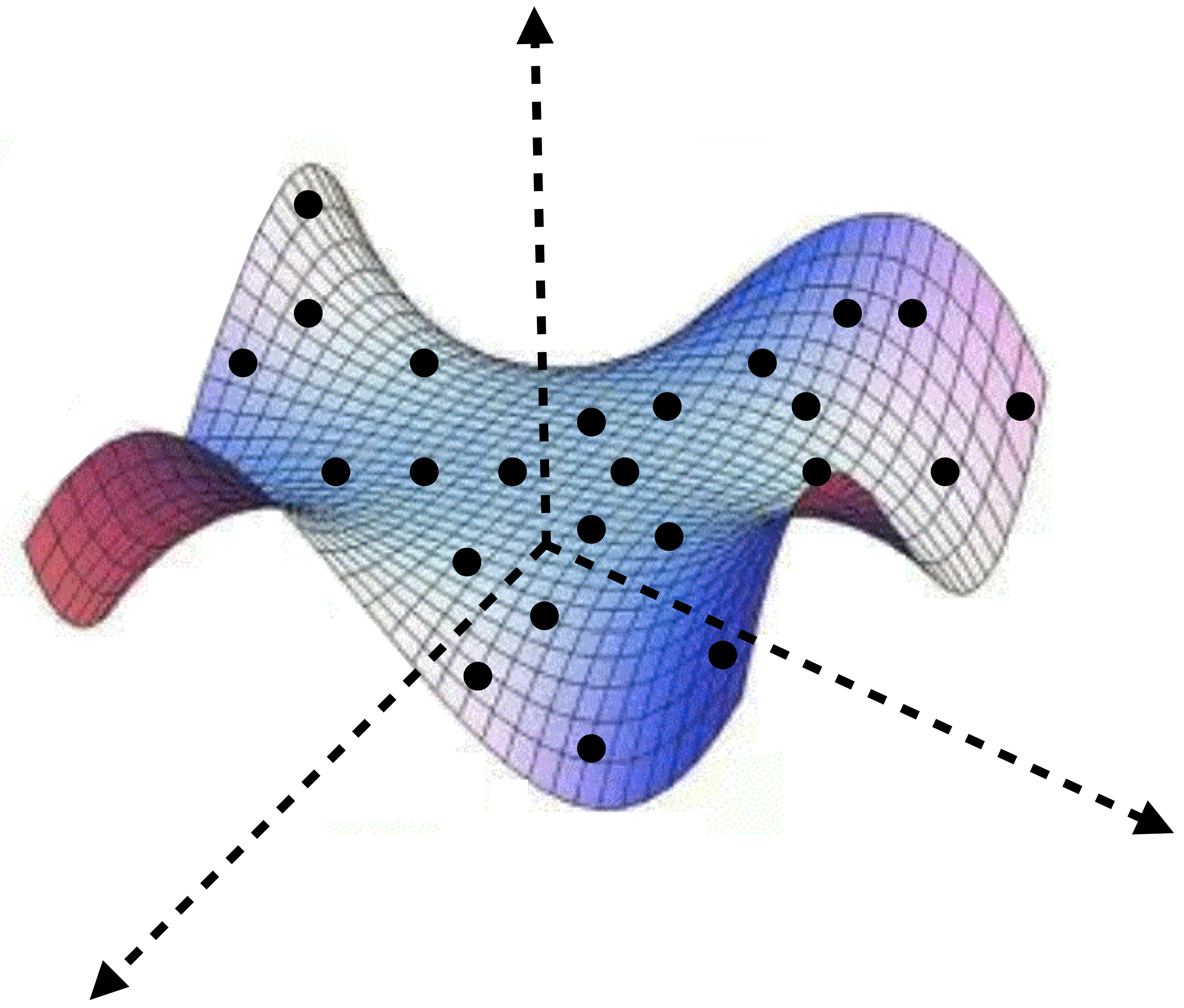
  $g_s : \underset{\text{low-dim}}{\mathbb{R}^k} \to \underset{\text{high-dim}}{\mathbb{R}^n}$

- g_a, g_s complex and unknown

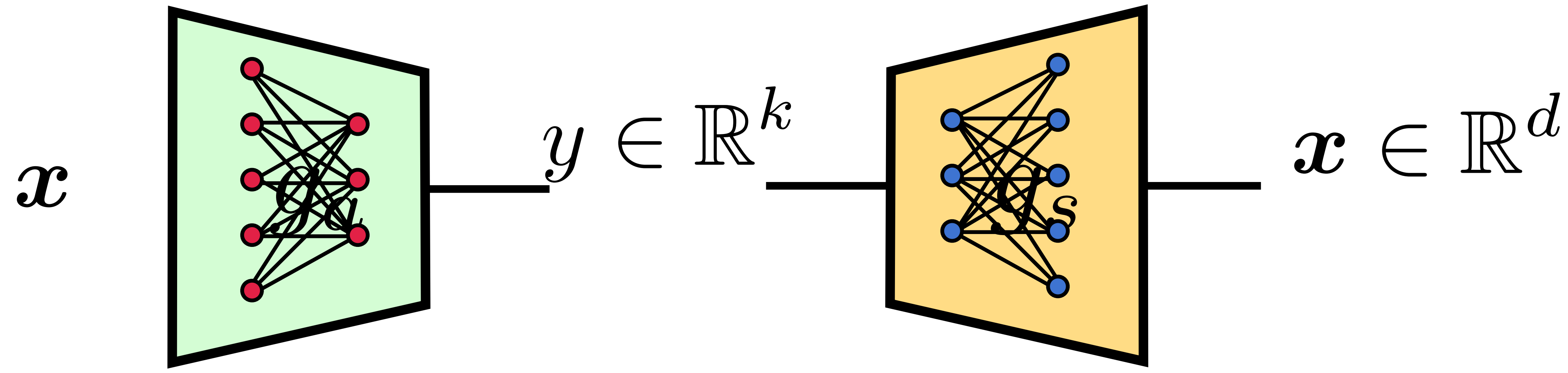


How is Learning Useful?

- Optimal schemes from information theory have exponential complexity in dimension
- Data is nominally high dimensional, but intrinsically is of much lower dimension

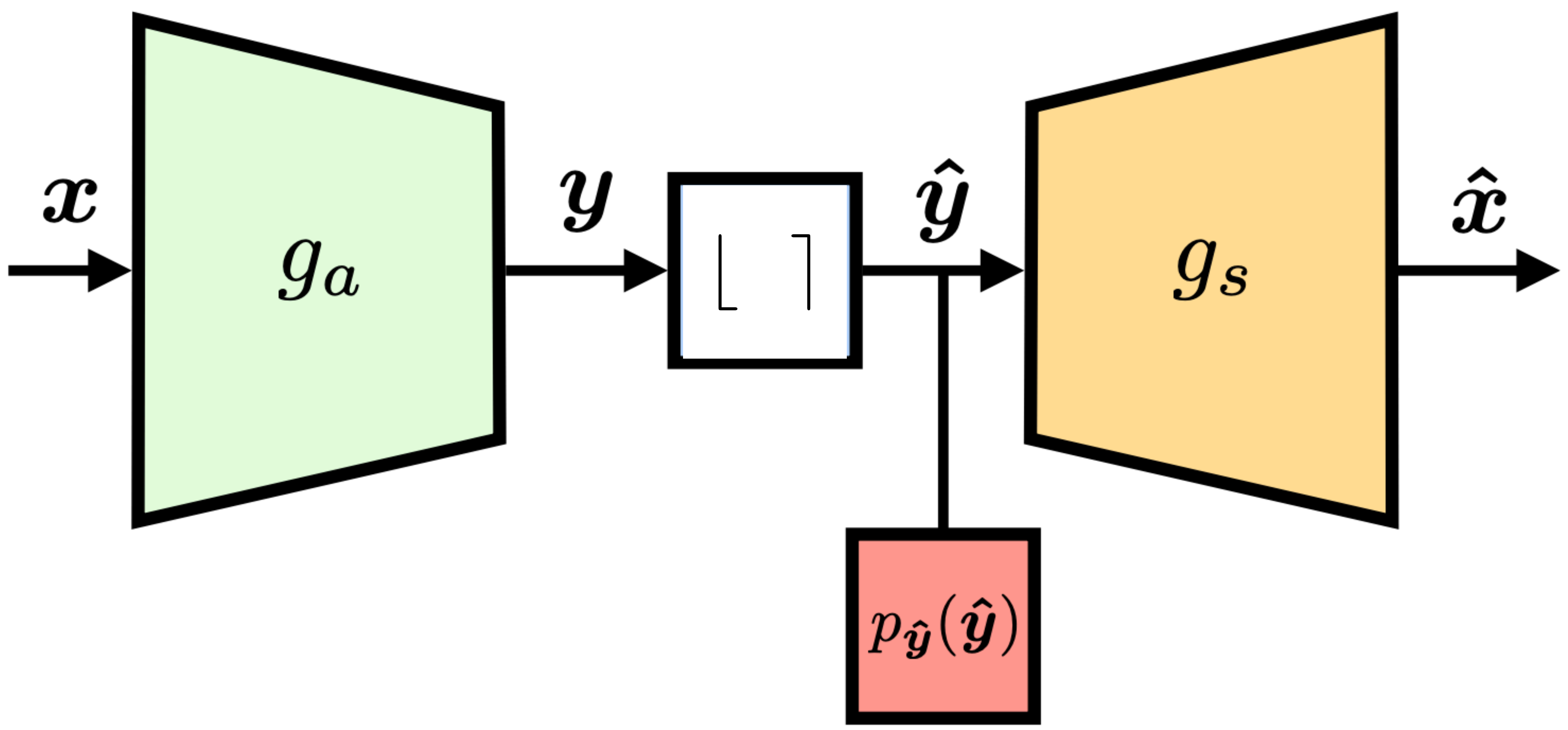


- the geometry:
  high-dim low-dim
  $g_a : \mathbb{R}^n \to \mathbb{R}^k$
  $g_s : \mathbb{R}^k \to \mathbb{R}^n$
  low-dim high-dim
- g_a, g_s complex and unknown
- learn it from data!



Neural Compression

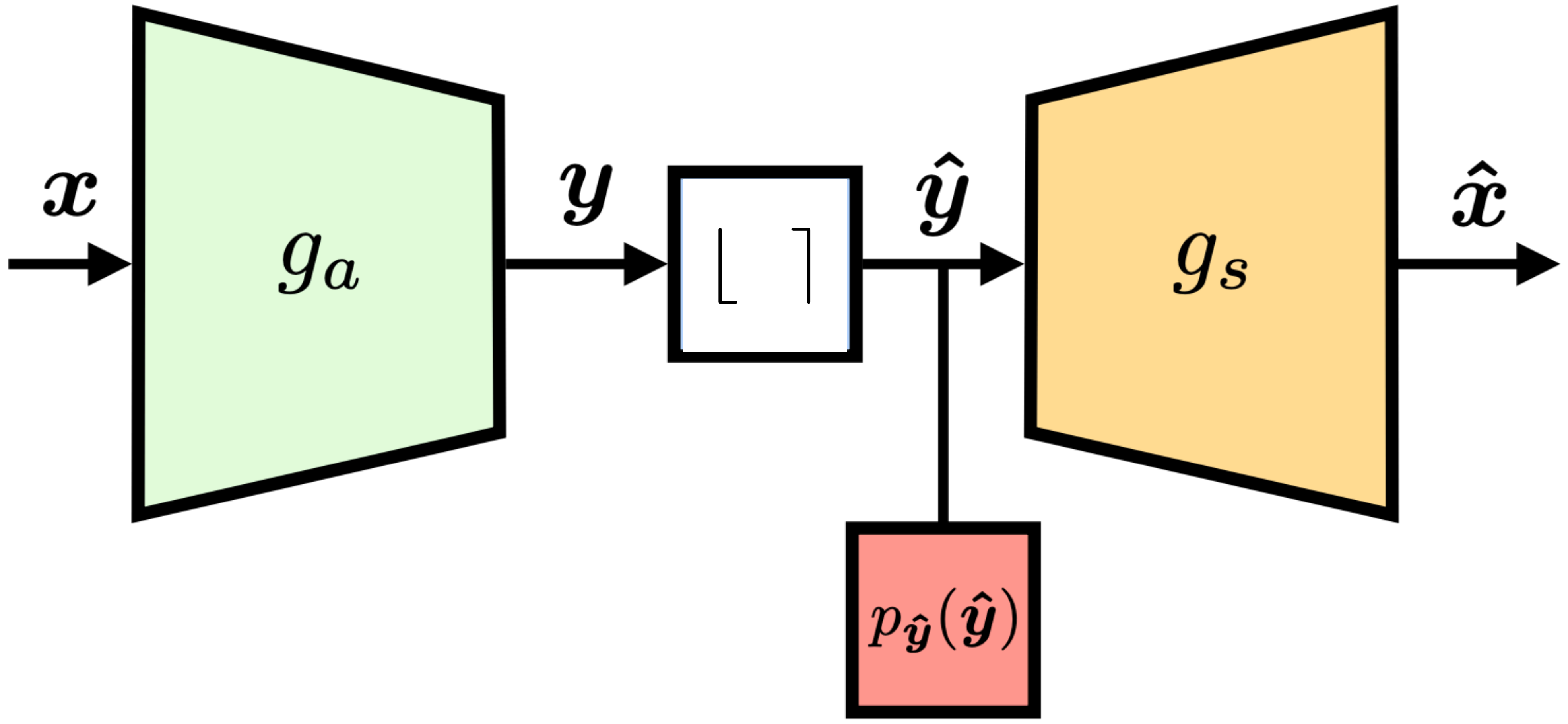
- Nonlinear Transform Coding (NTC)
- Transform x to y
- y is rounded to $\hat{y}$ entry-wise
- $\hat{y}$ is encoded under model $p_{\hat{y}}$ (also learned)
- Reconstruction $\hat{x}$ is transformed from $\hat{y}$
- Objective: $\min_{g_a, g_s, p_{\hat{y}}} \mathbb{E}_{\boldsymbol{x}} [-\log p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}})] + \lambda \cdot \mathbb{E}_{\boldsymbol{x}}[\mathsf{d}(\boldsymbol{x}, \hat{\boldsymbol{x}})]$



[Theis et al ’17] [Agustsson et al ‘17]
[Ballé et al ’17] [Minnen et al ’18]

(rate/distortion tradeoff)

Neural Compression



- Nonlinear Transform Coding (NTC)
- Transform $\boldsymbol{x}$ to $\boldsymbol{y}$
- $\boldsymbol{y}$ is rounded to $\hat{\boldsymbol{y}}$ entry-wise
- $\hat{\boldsymbol{y}}$ is encoded under model $p_{\hat{\boldsymbol{y}}}$ (also learned)
- Reconstruction $\hat{\boldsymbol{x}}$ is transformed from $\hat{\boldsymbol{y}}$
- Objective: $\min_{g_a, g_s, p_{\hat{\boldsymbol{y}}}} \mathbb{E}_{\boldsymbol{x}} \left[-\log p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}}) \right] + \lambda \cdot \mathbb{E}_{\boldsymbol{x}} [\mathsf{d}(\boldsymbol{x}, \hat{\boldsymbol{x}})]$

[Theis et al '17] [Agustsson et al '17] [Ballé et al '17] [Minnen et al '18]

(rate/distortion tradeoff)

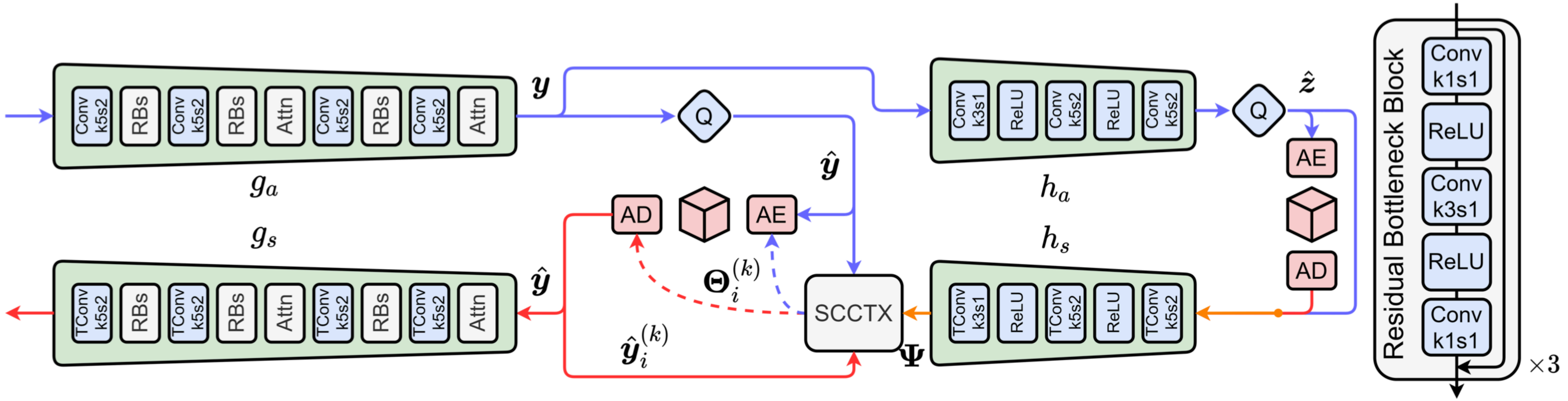
$$\min \mathbb{E}[-\log p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}})] + \lambda_1 \mathbb{E}[d(\boldsymbol{x}, \hat{\boldsymbol{x}})] + \lambda_2 \delta(P_{\boldsymbol{x}}, P_{\hat{\boldsymbol{x}}})$$

(rate/distortion/perception tradeoff)

[Mentzer '22] [Muckley et al '23] [Agustsson et al '23]

Recent Architectures

- Recent architectures involve sophisticated transform + entropy model design [1, 2, 3]
- Training: noisy proxy $\lfloor g_a(\boldsymbol{x}) \rceil \rightarrow g_a(\boldsymbol{x}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{Unif}([-0.5, 0.5)^d)$
- Entropy model $p_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) = \left[\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) * U(-0.5, 0.5)\right](\hat{\boldsymbol{y}})$
- Complex channel-spatial dependencies within $\hat{\boldsymbol{y}}$



ELIC [1]

[1] He, Dailan, et al. "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding." CVPR 2022.
[2] He, Dailan, et al. "Po-elic: Perception-oriented efficient learned image coding" CVPR 2022.
[3] M. Muckley et al. "Improving statistical fidelity for neural image compression with implicit local likelihood models." ICML 2023.

Fundamental Questions

- Are learning-based compressors such as NTC information-theoretically optimal?
  - Some look at stylized sources with intrinsic dimension one
    [Wagner&Ballé ’21], [Bhadane et al ’22], [Ozyilkan et al ‘24]
  - Some compute bounds on the RD function of real-world sources and show that there is a gap
    [Lei, Hassani, SB ’22], [Yang&Mandt ’22]

- Can we design practical compressors informed by information theoretic designs?

Outline

Outline

- Sub-optimality of NTC for Gaussian sources

Outline

- Sub-optimality of NTC for Gaussian sources
- Lattice Transform Coding (LTC) for RD

Outline

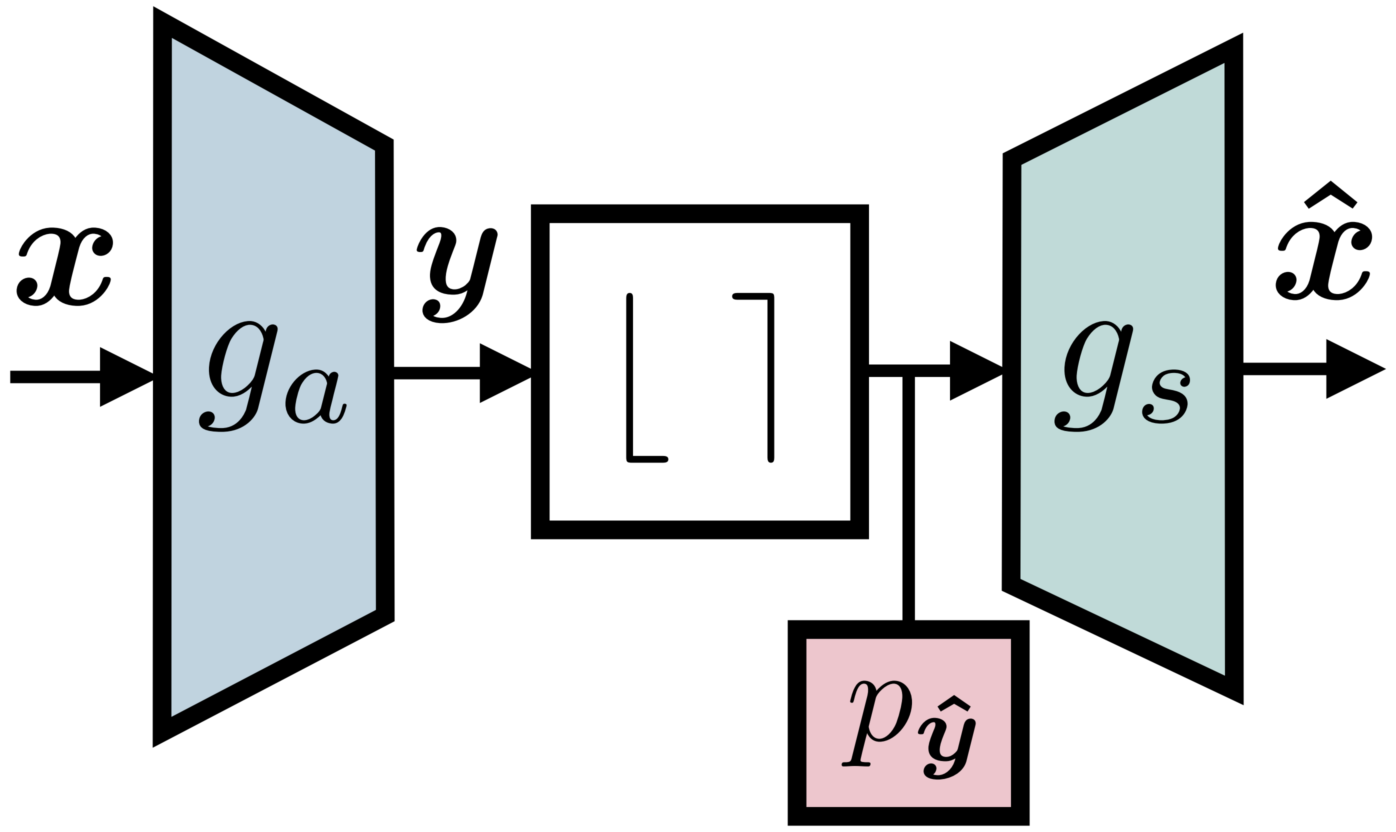
- Sub-optimality of NTC for Gaussian sources
- Lattice Transform Coding (LTC) for RD
- LTC with Dithering for RDP

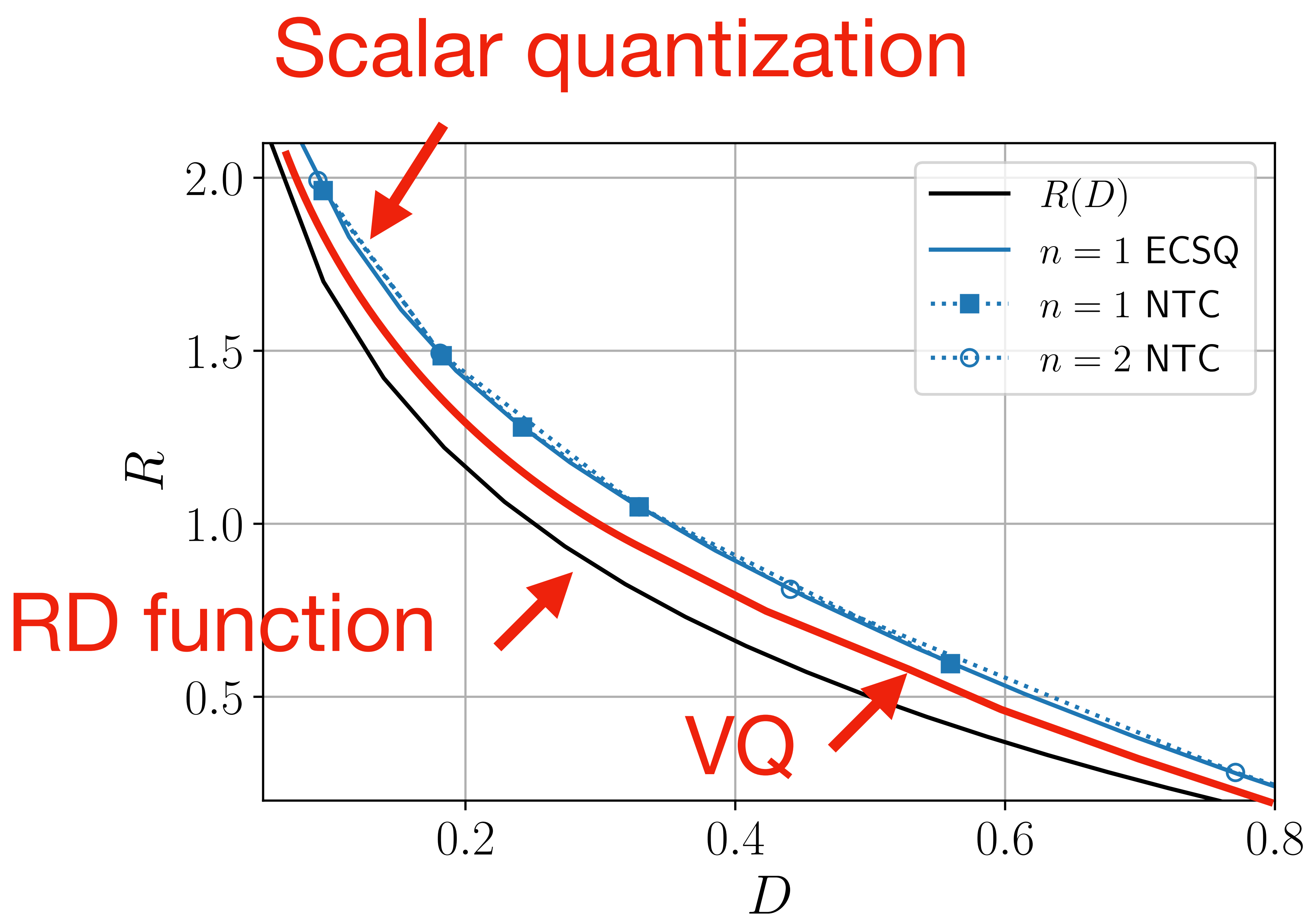
Outline

- Sub-optimality of NTC for Gaussian sources
- Lattice Transform Coding (LTC) for RD
- LTC with Dithering for RDP
- Simulation Results

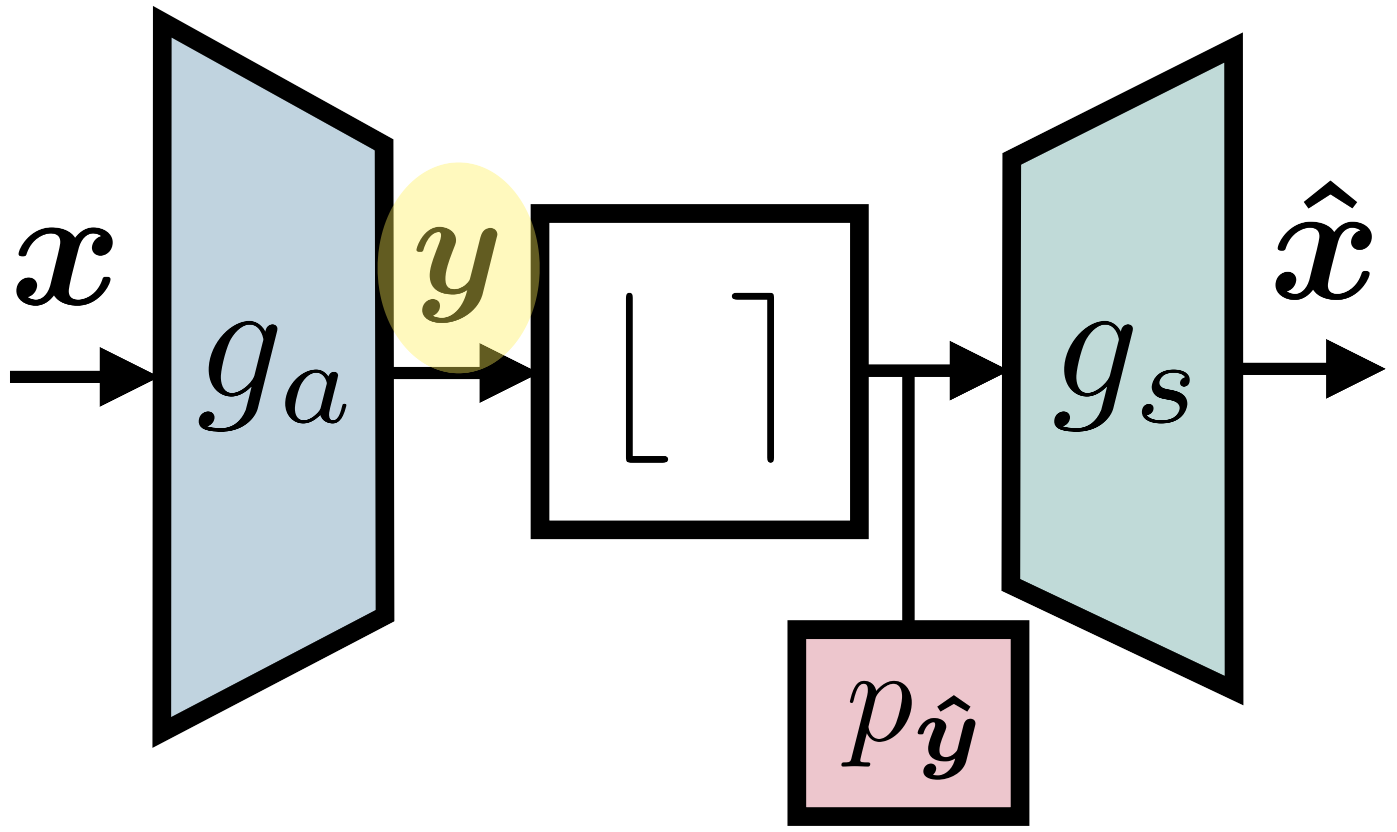
NTC for i.i.d. Gaussian Source

- Source: $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n), \quad \boldsymbol{x}_i \sim \mathcal{N}(0,1)$
- Consider $n = 1,2,\ldots$
- NTC does not outperform scalar quantization with increasing n





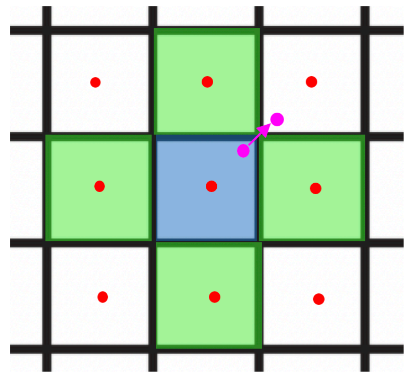
Lattice Packings



- In NTC, the latent vector is rounded element-wise

  Equivalent to the integer lattice

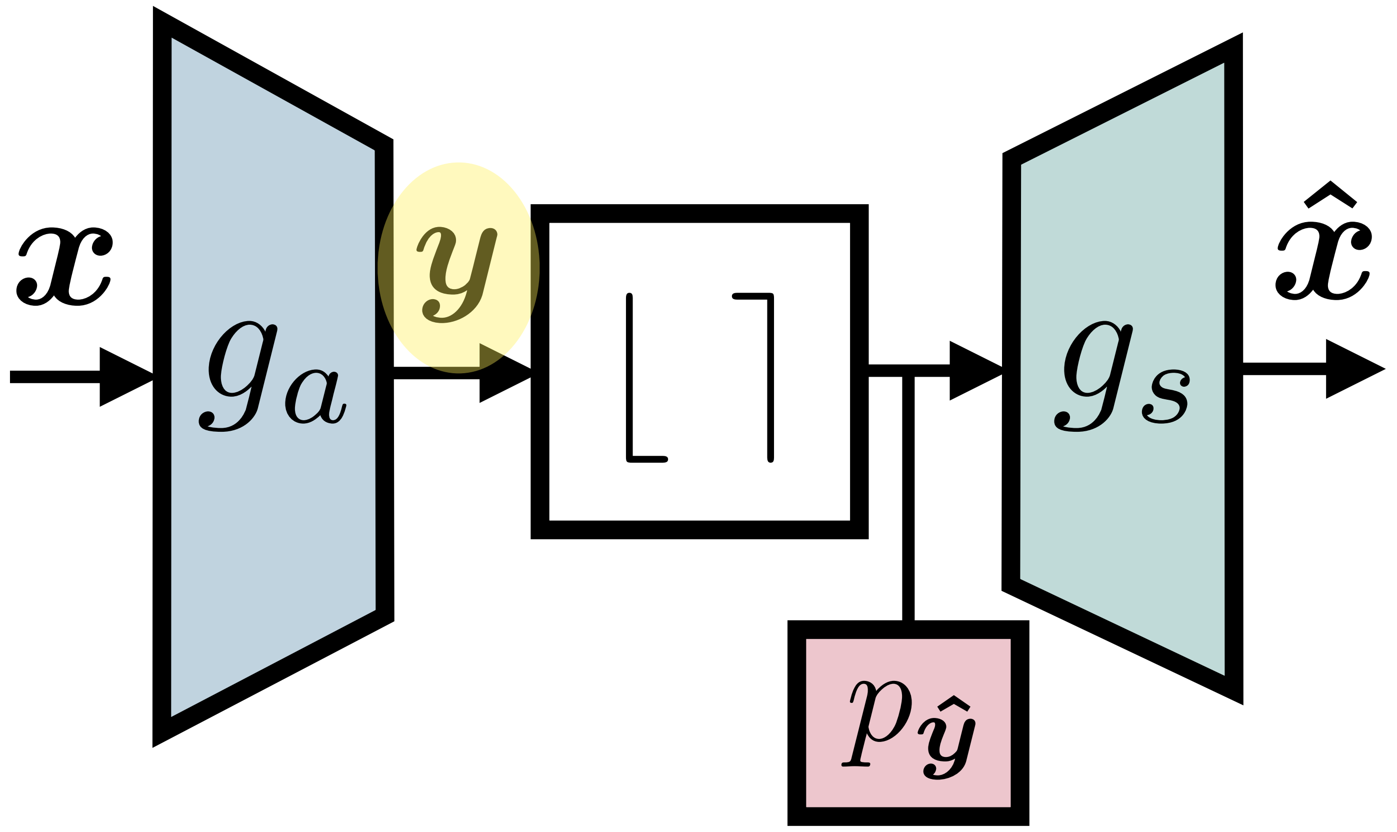
  Not the most efficient in packing the space



Integer Lattice

[1] Lei, Eric, Hamed Hassani, and Shirin Saeedi Bidokhti. "Approaching Rate-Distortion Limits in Neural Compression with Lattice Transform Coding." *ICLR. 2025*

Lattice Packings



- In NTC, the latent vector is rounded element-wise

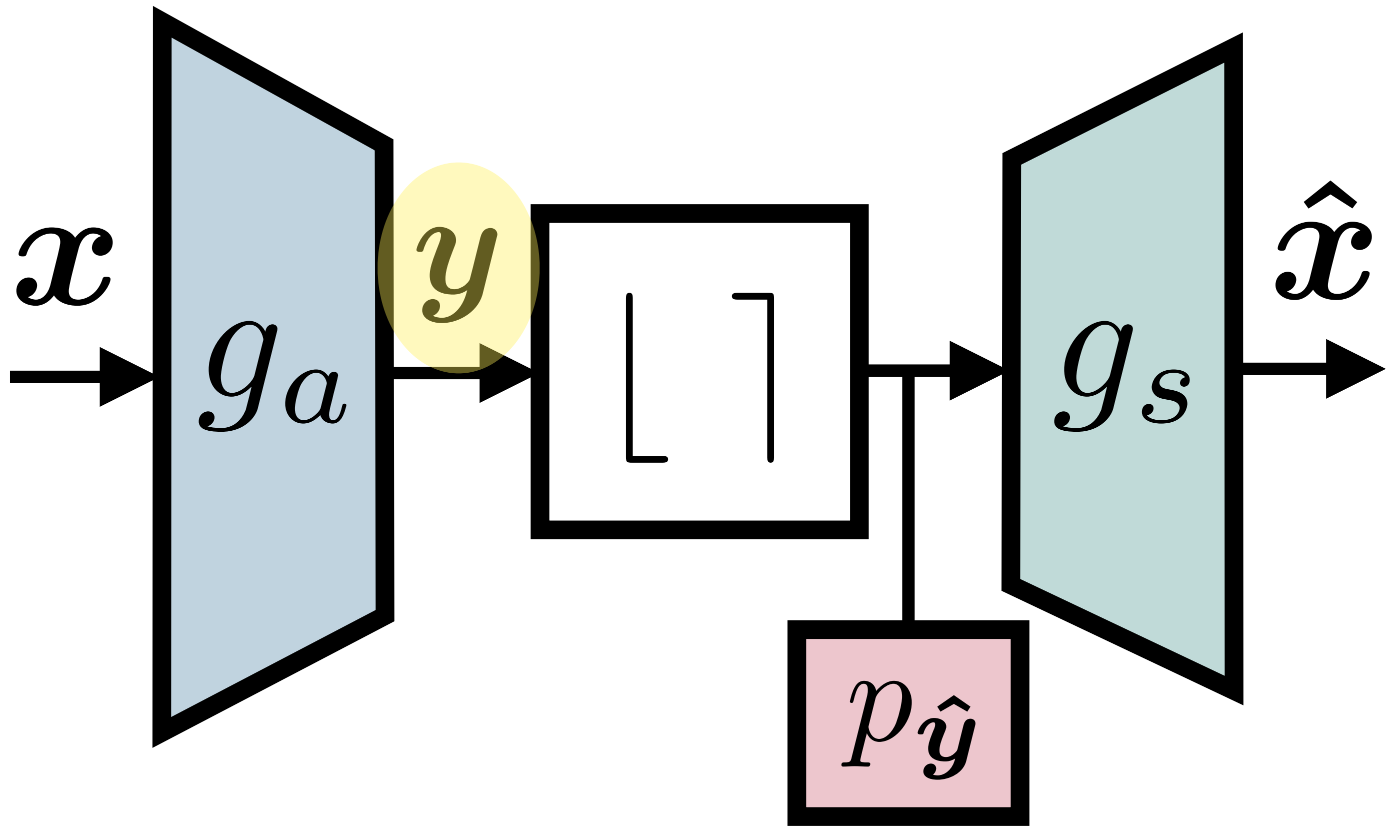
  Equivalent to the integer lattice

  Not the most efficient in packing the space

Integer Lattice   Hexagonal Lattice

[1] Lei, Eric, Hamed Hassani, and Shirin Saeedi Bidokhti. "Approaching Rate-Distortion Limits in Neural Compression with Lattice Transform Coding." *ICLR. 2025*

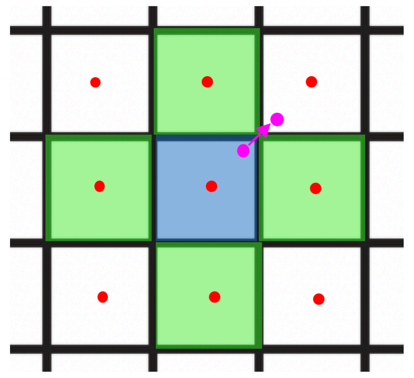
Lattice Packings



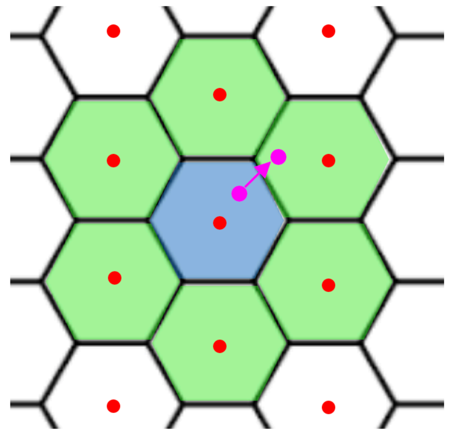
- In NTC, the latent vector is rounded element-wise

  Equivalent to the integer lattice

  Not the most efficient in packing the space



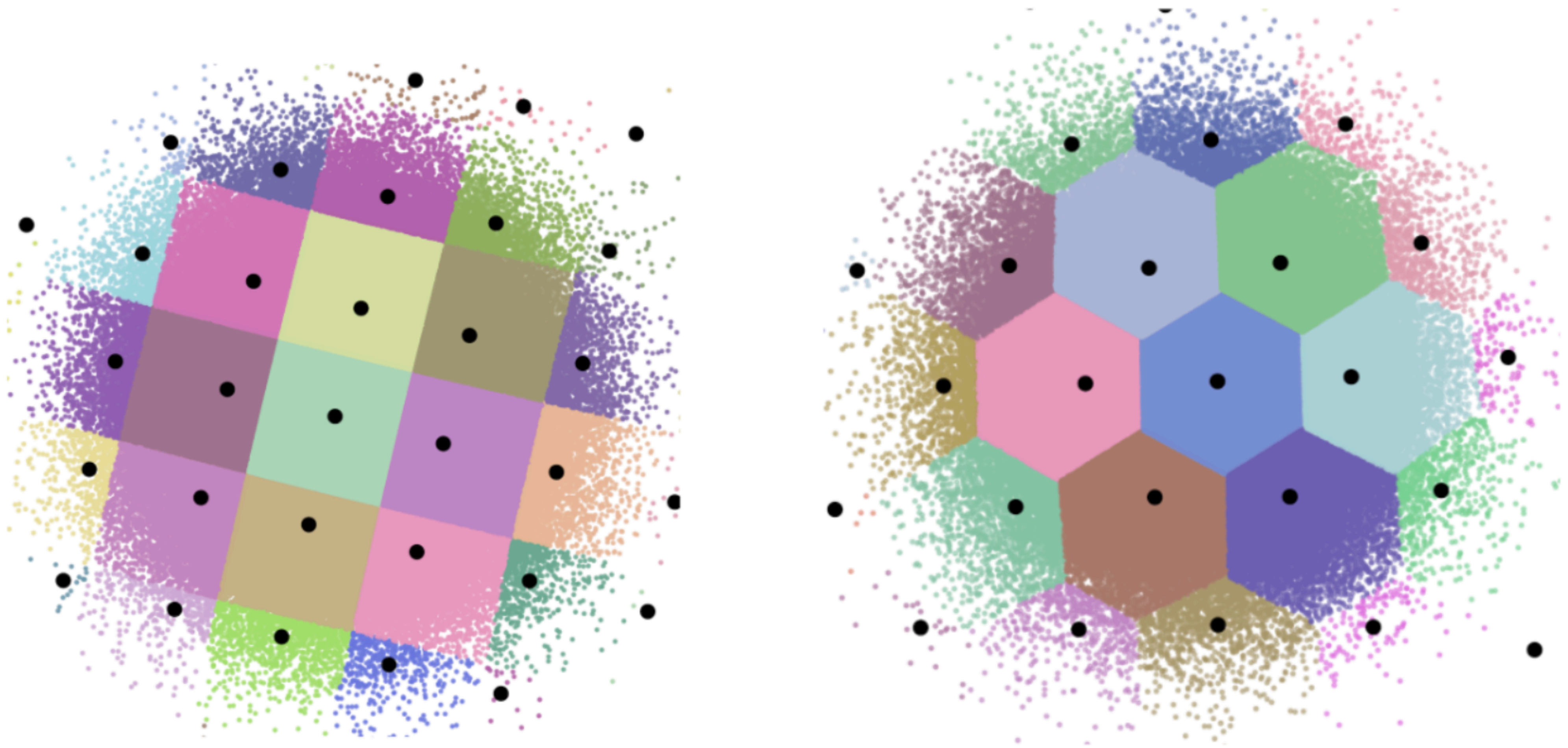
Integer Lattice



Hexagonal Lattice

- g_a, g_s fail to map square regions to hexagons

  Increasing depth/width does not help



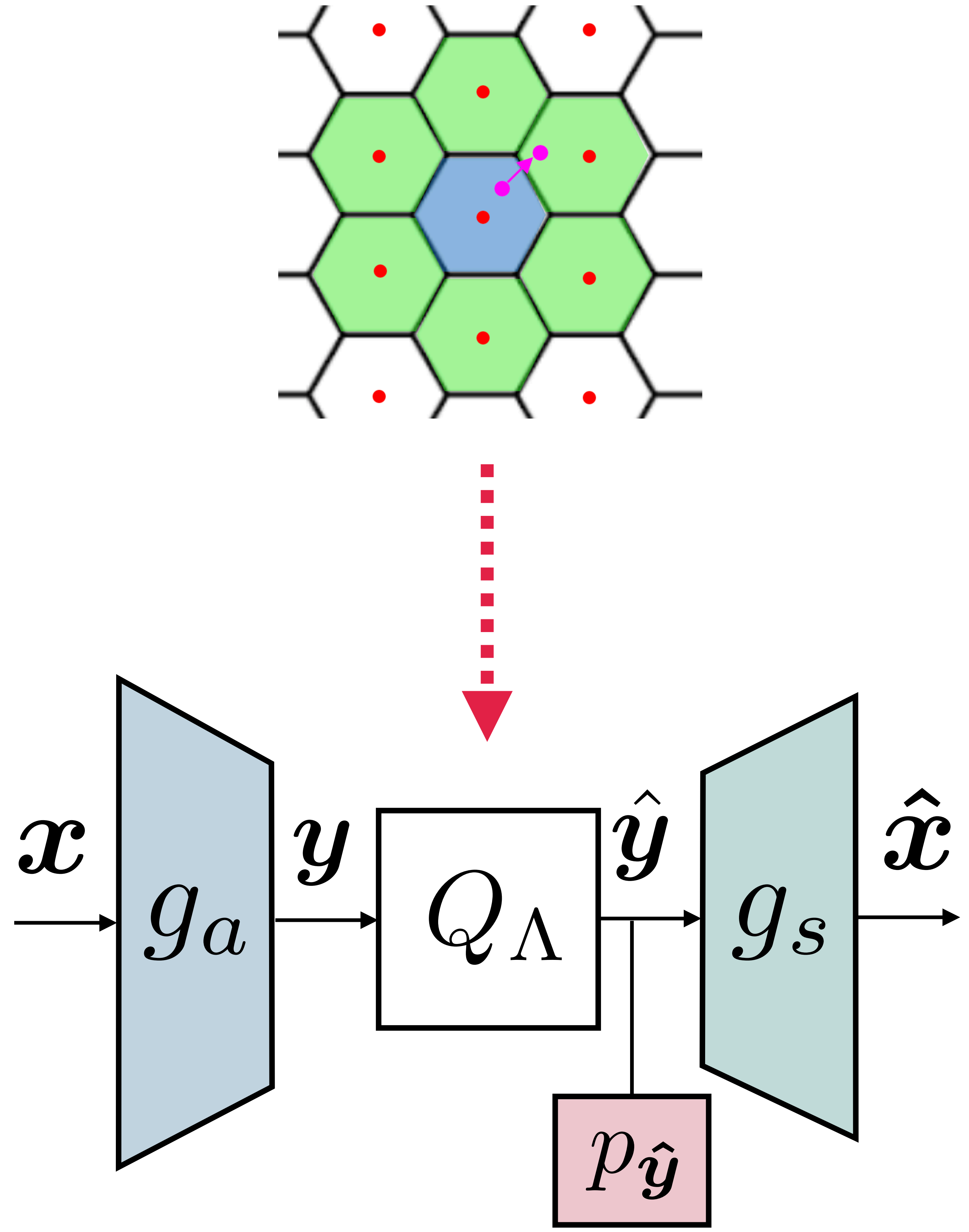
NTC

Optimal VQ

Quantization Regions (2-d)

[1] Lei, Eric, Hamed Hassani, and Shirin Saeedi Bidokhti. "Approaching Rate-Distortion Limits in Neural Compression with Lattice Transform Coding." *ICLR. 2025*

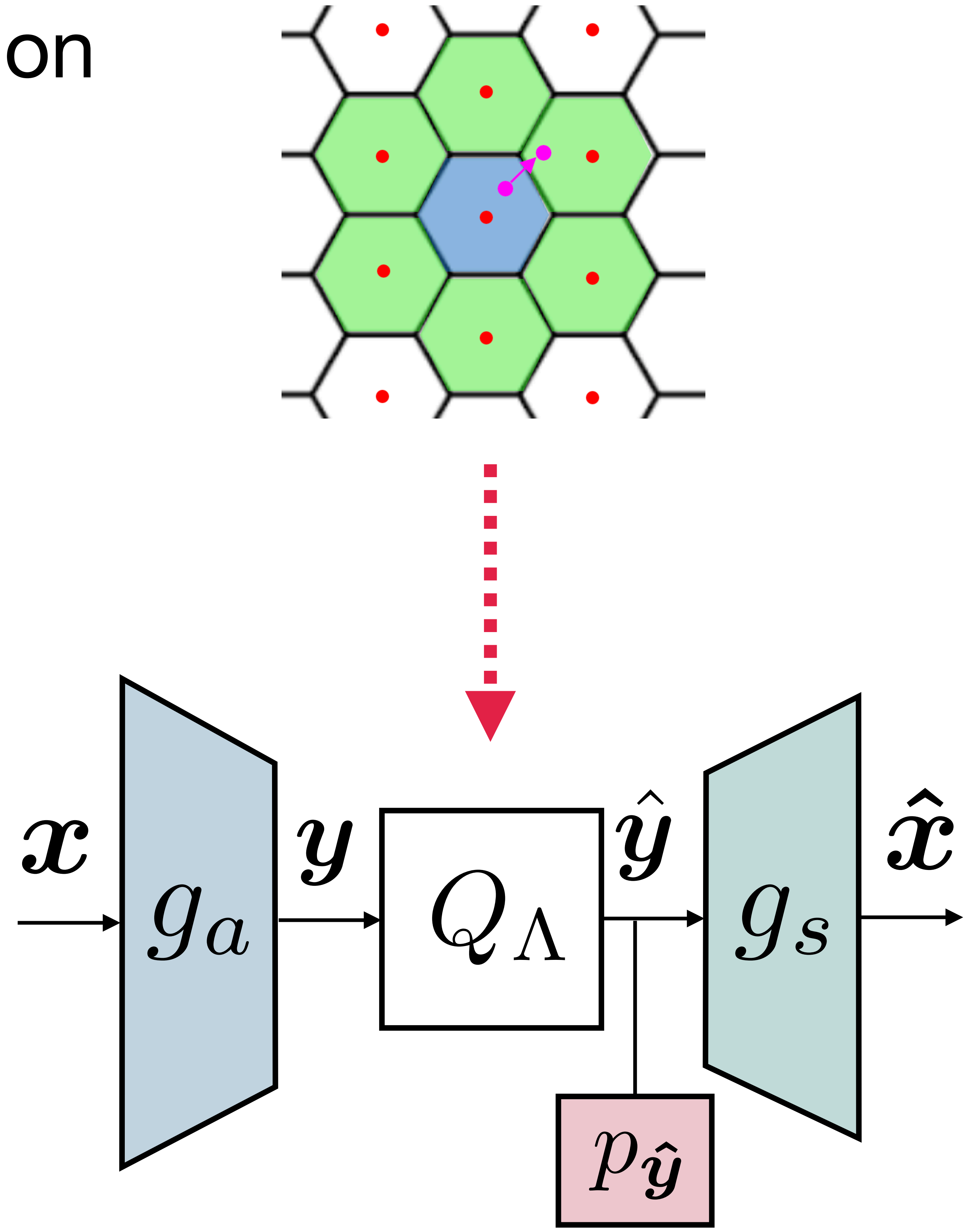
Lattice Quantization in the Latent Space



[1] Lei, Eric, Hamed Hassani, and Shirin Saeedi Bidokhti. "Approaching Rate-Distortion Limits in Neural Compression with Lattice Transform Coding." *ICLR. 2025*

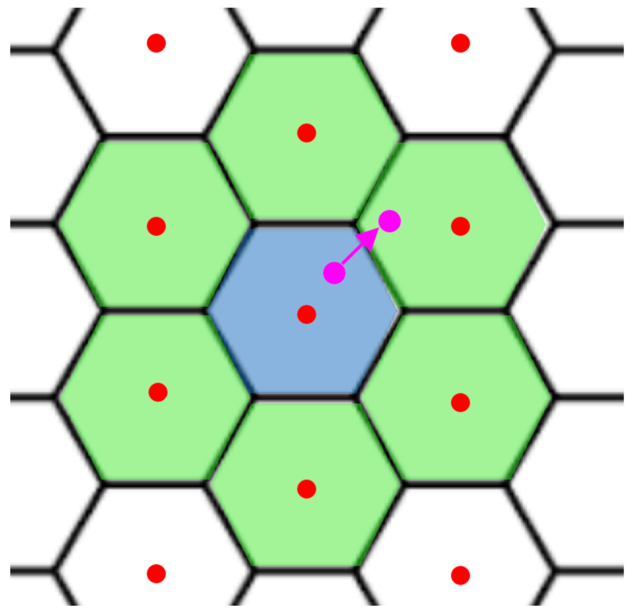
Lattice Quantization in the Latent Space

- Idea: Replace the integer rounding, with lattice quantization

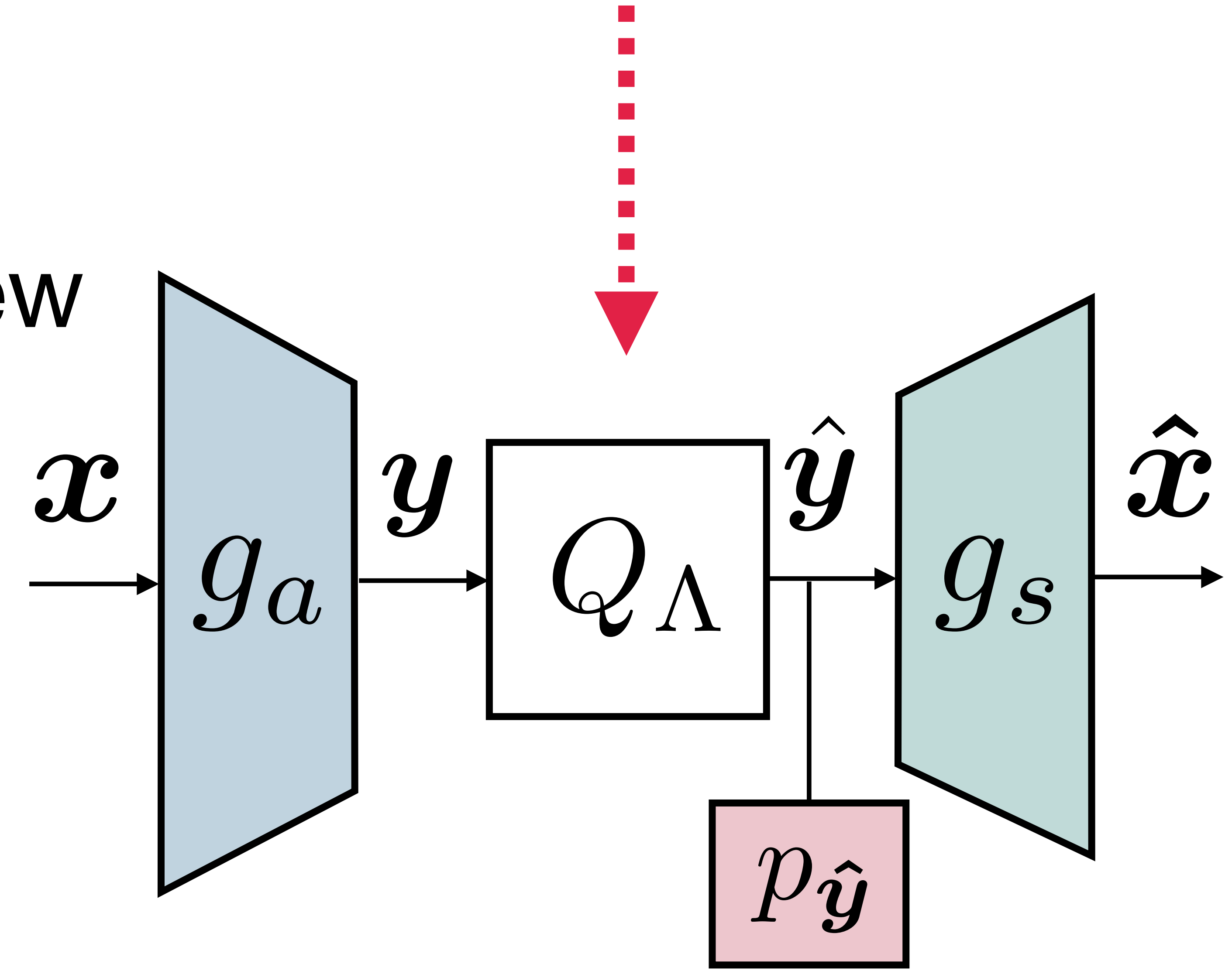


[1] Lei, Eric, Hamed Hassani, and Shirin Saeedi Bidokhti. "Approaching Rate-Distortion Limits in Neural Compression with Lattice Transform Coding." *ICLR. 2025*

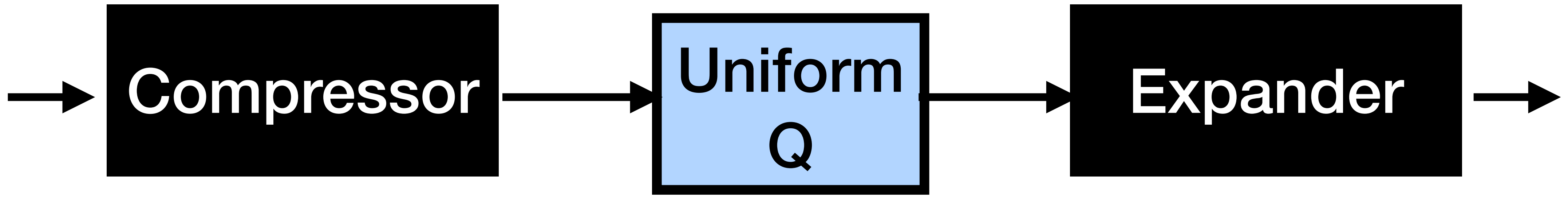
Lattice Quantization in the Latent Space



- Idea: Replace the integer rounding, with lattice quantization



- Connection to companding results [Gersho 1979; Bucklew 1981; Bucklew 1983; Linder-Zamir-Zeger 1999]

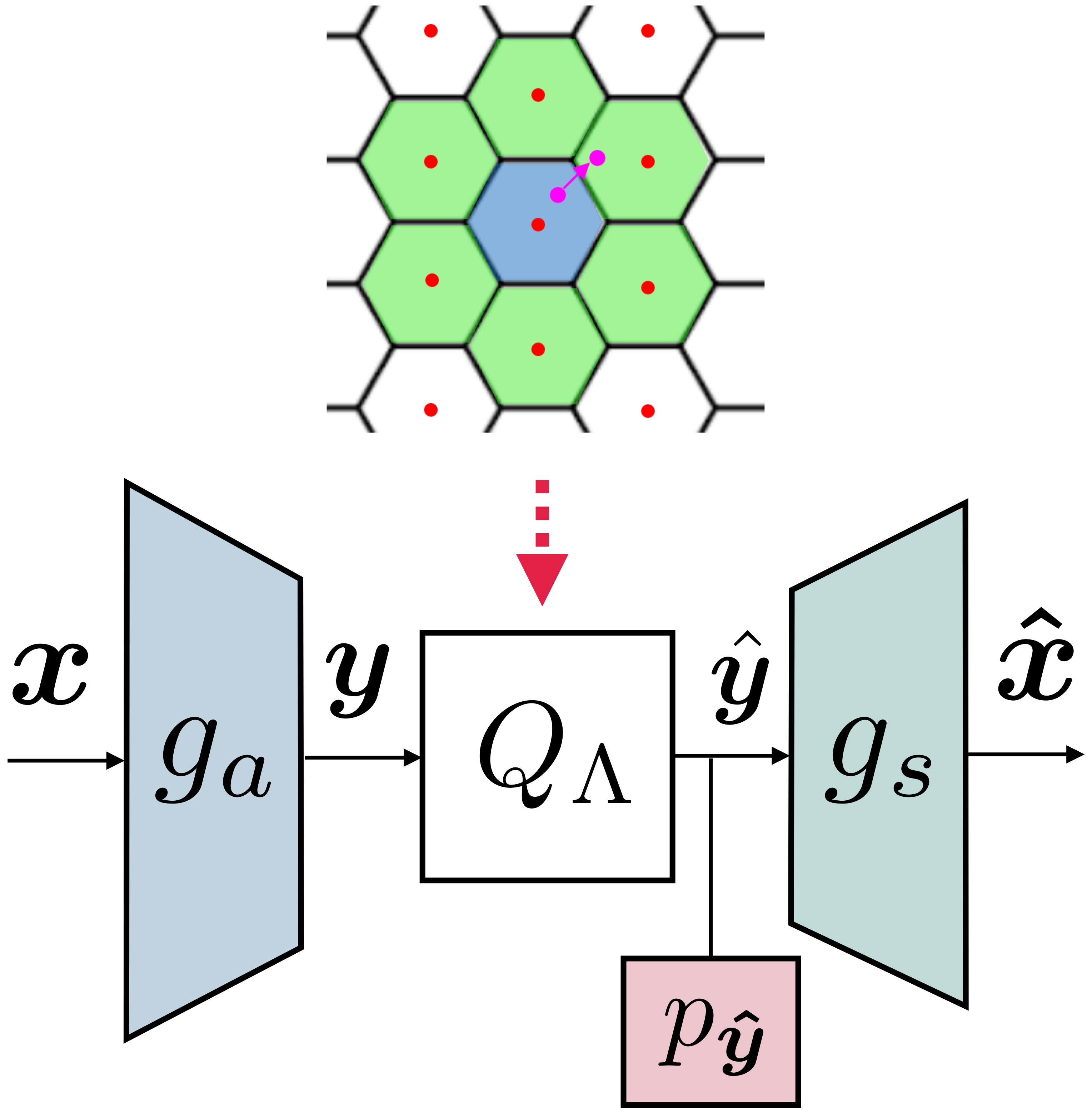


- Asymptotically RD- optimal for Gaussian sources

[1] Lei, Eric, Hamed Hassani, and Shirin Saeedi Bidokhti. "Approaching Rate-Distortion Limits in Neural Compression with Lattice Transform Coding." *ICLR. 2025*

Lattice Transform Coding

- Lattice Transform Coding (LTC)
- Transform x to y
- y is lattice-quantized to $\hat{y}$
- $\hat{y}$ is encoded under model $p_{\hat{y}}$ (also learned)
- Reconstruction $\hat{x}$ is transformed from $\hat{y}$
- Objective: $\min_{g_a, g_s, p_{\hat{y}}} \mathbb{E}_{\boldsymbol{x}} [-\log p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}})] + \lambda \cdot \mathbb{E}_{\boldsymbol{x}}[\mathsf{d}(\boldsymbol{x}, \hat{\boldsymbol{x}})]$



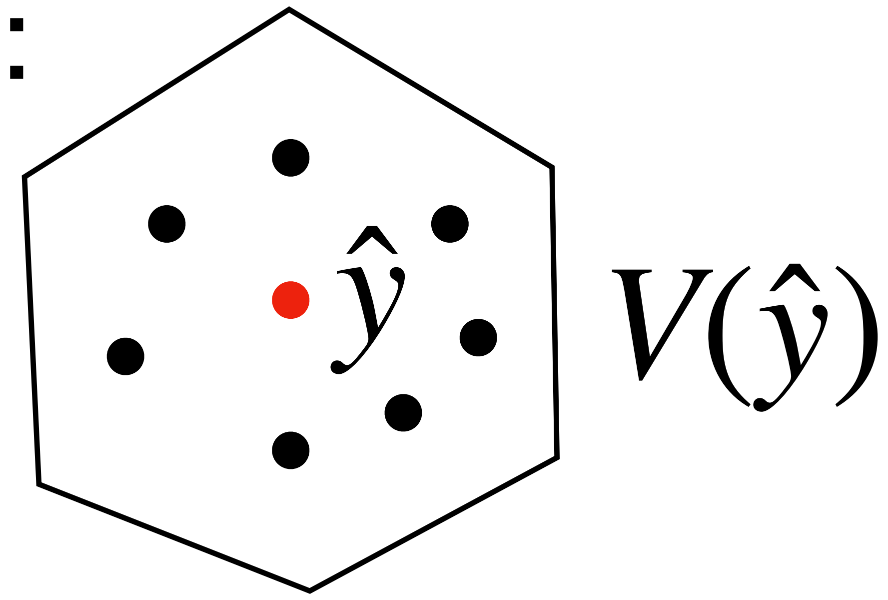
- Using lattices requires new methods to optimizing the objective…

Computing the Rate Term

- Objective: $\min_{g_a, g_s, p_{\hat{\boldsymbol{y}}}} \mathbb{E}_{\boldsymbol{x}}\left[-\log p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}})\right] + \lambda \cdot \mathbb{E}_{\boldsymbol{x}}[d(\boldsymbol{x}, \hat{\boldsymbol{x}})]$

- PMF on centers $\hat{y}$ defined by integrating PDF $p_{\boldsymbol{y}}(\boldsymbol{y})$ over latent space:

$$p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}}) = \int_{V(\hat{\boldsymbol{y}})} p_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y}$$



- In NTC, lattice cell $V(\hat{y})$ is a square— easy to integrate
- For a lattice, $V(\hat{y})$ is no longer square— difficult to integrate!
- Instead, we integrate using Monte-Carlo: $p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}}) = \mathbb{E}_{\boldsymbol{u}' \sim \mathrm{Unif}(V(\mathbf{0}))}[p_{\boldsymbol{y}}(\hat{\boldsymbol{y}} + \boldsymbol{u}')]$

The Choice of the Lattice Λ

- Larger lattice dimension n —> improved packing efficiency
- Complexity— finding closest lattice vector
- Densest lattices for $n \leq 24$ with low complexity
    - $n = 2$ Hexagonal lattice
    - $n = 4$: D_n^* lattice
    - $n = 8$: E_8 (Gosset) lattice
    - $n = 16$: Λ_{16} (Barnes-Wall) lattice
    - $n = 24$: Λ_{24} (Leech) lattice

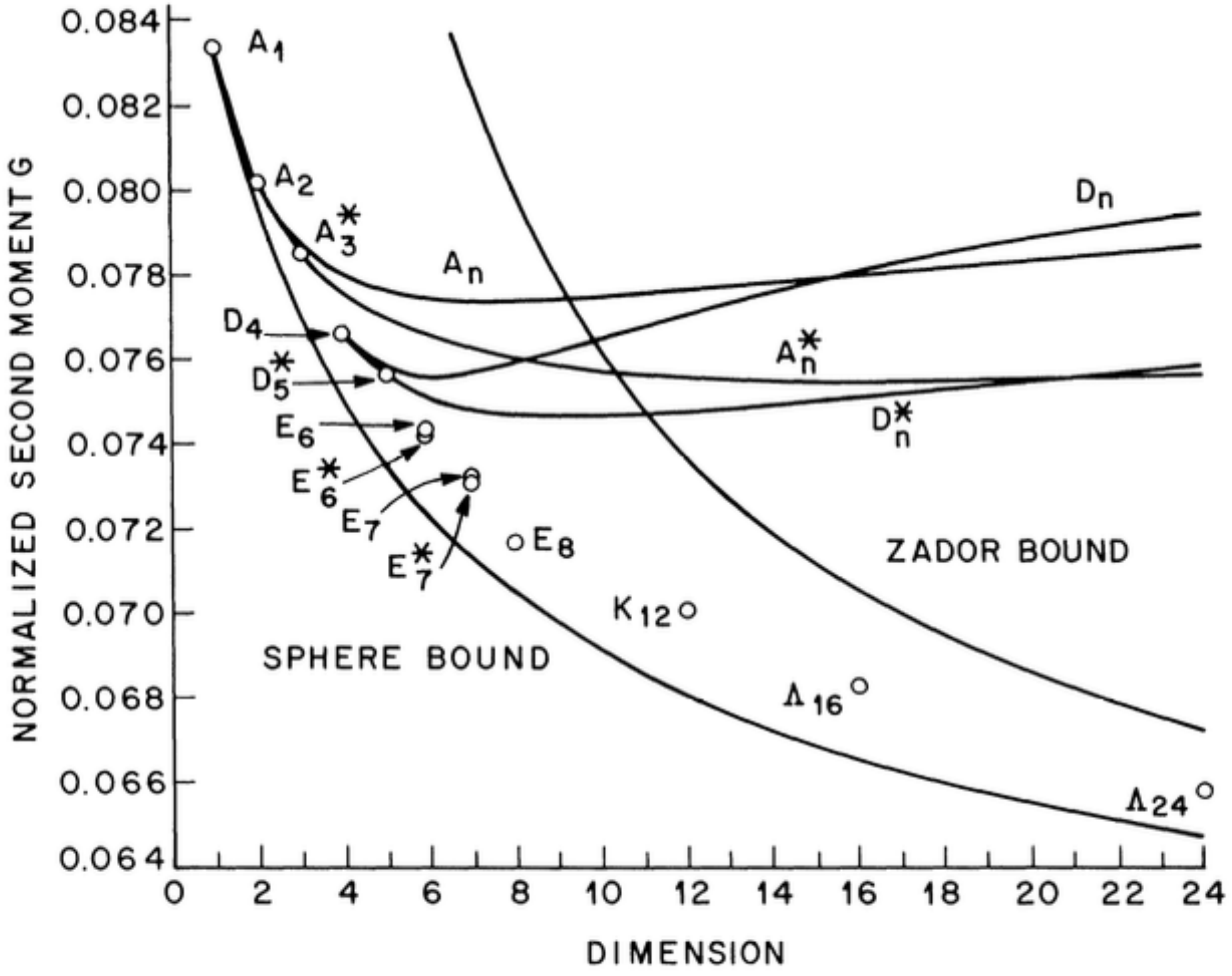


FIG. 2. *Normalized second moment G for various lattices, and the Zador and sphere bounds. It is known that the best quantizers must lie between the two bounds.*

[Conway and Sloane, 1984]

LTC for i.i.d. Gaussian Source

- Source: $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n), \quad \boldsymbol{x}_i \sim \mathcal{N}(0, 1)$
- Consider $n = 2,\ 4,\ 8,\ 24$
- LTC performs close to VQ
  - Does not require exponential codebook search
- Approaches $R(D)$ lower bound

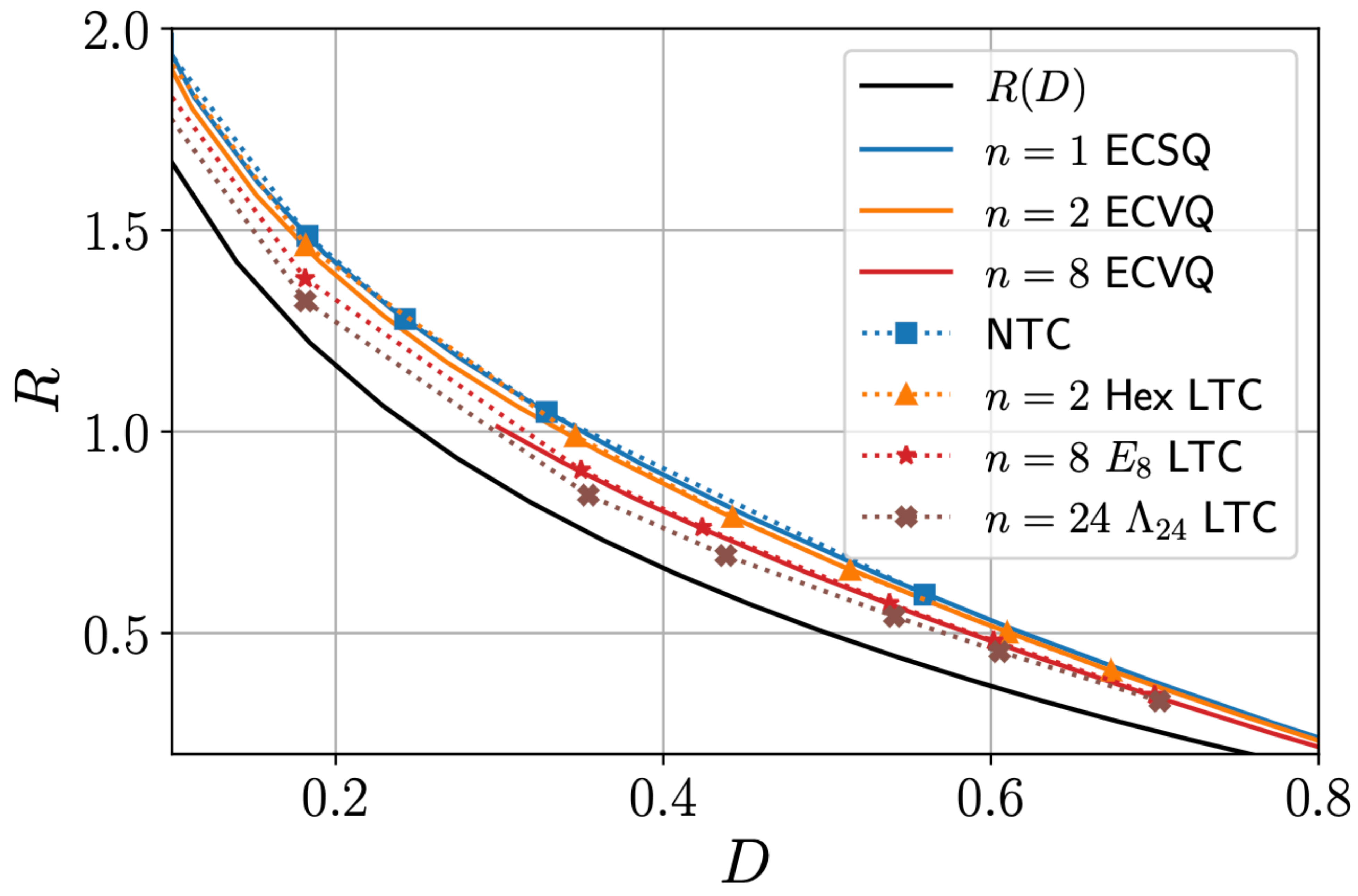
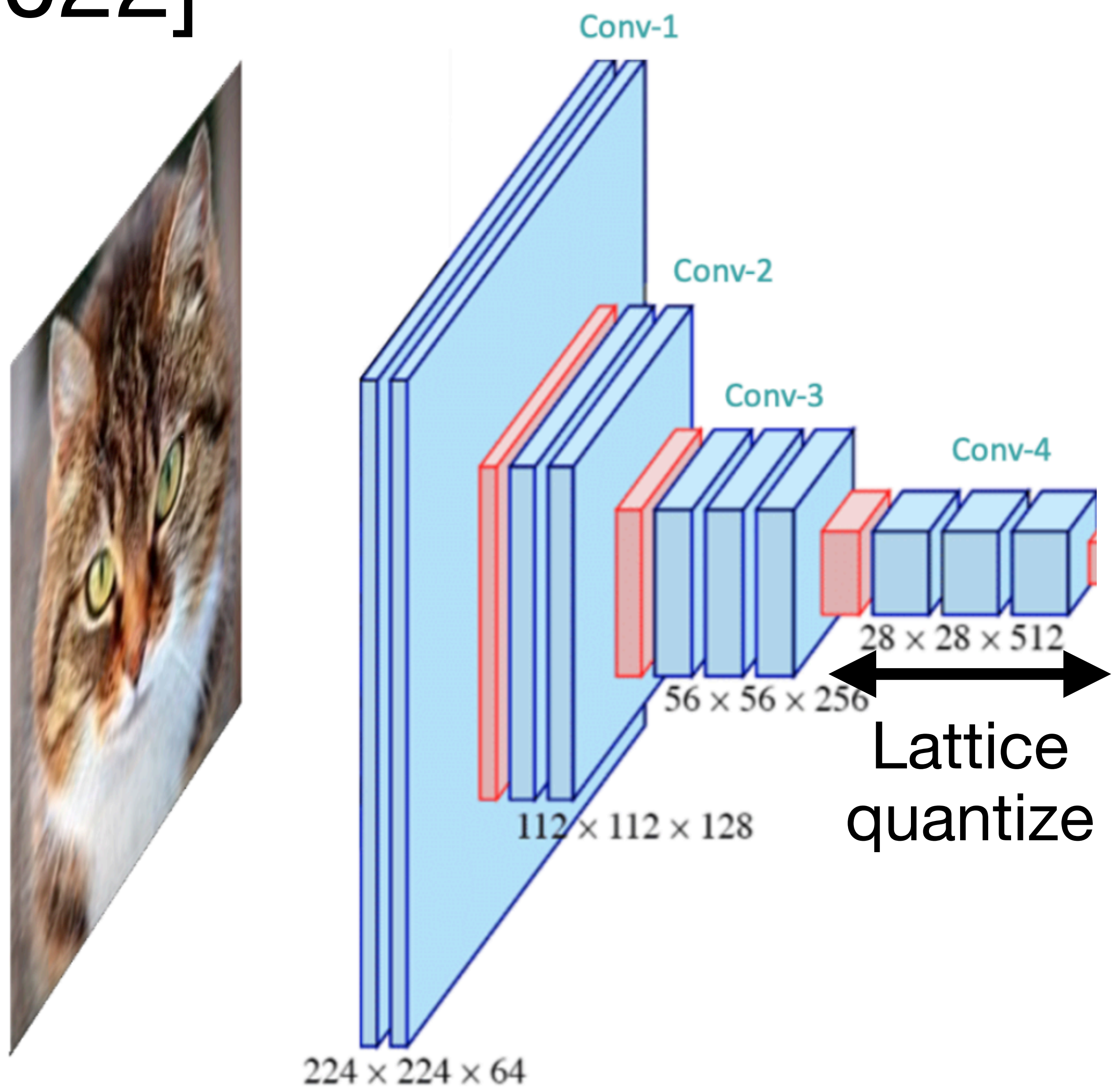
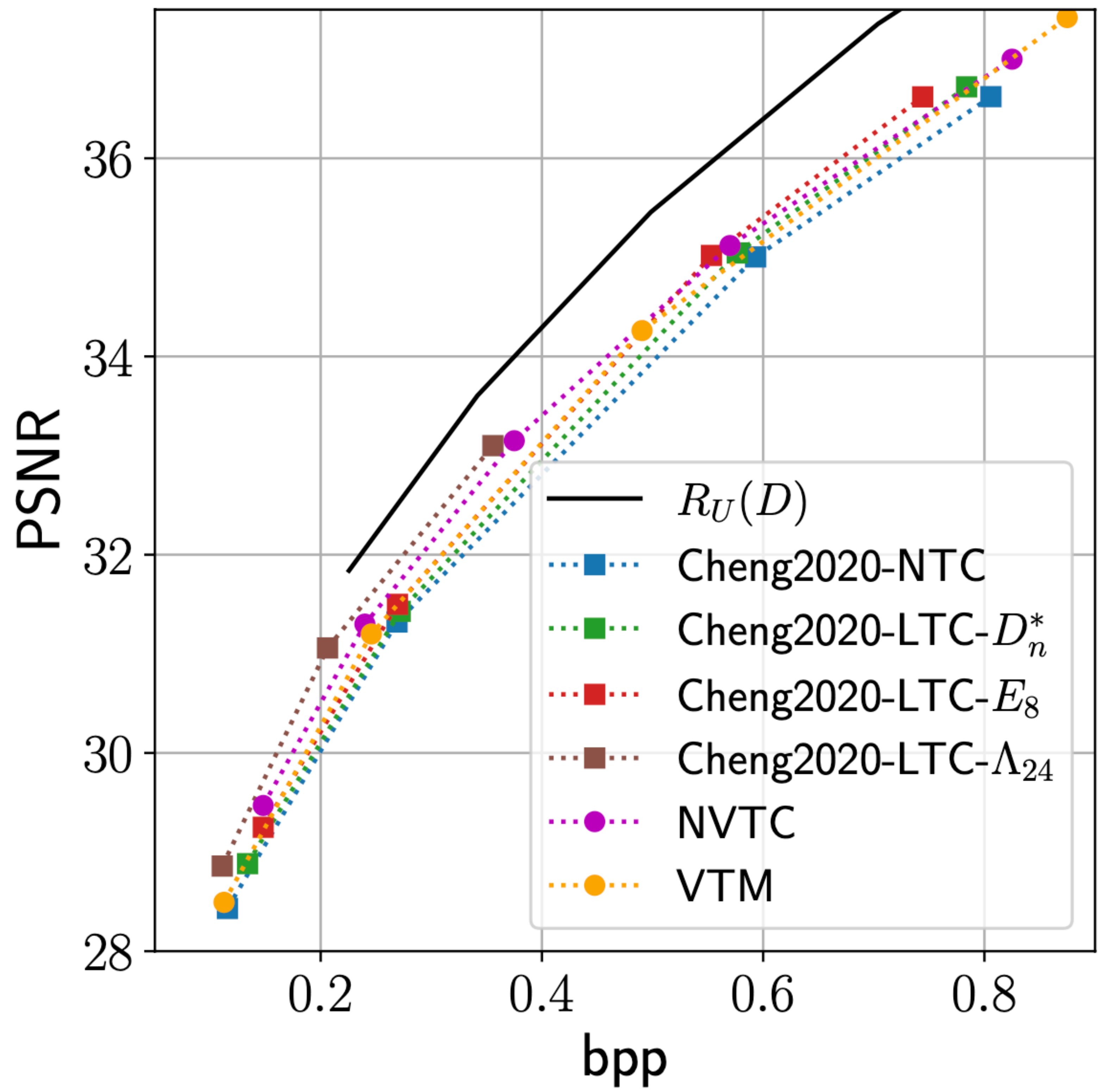


Image Compression

- Apply lattices along "channel" dimension of latent tensor
- Apply lattices product-wise
- Outperforms VTM and recent VQ-based codecs
- Approaches Kodak $R(D)$ bound from [Yang and Mandt, 2022]





Kodak evaluation dataset

So Far…

- Lattice transform coding (LTC), uses latent lattice quantization, and can recover VQ without exponential complexity

- Toward RDP …

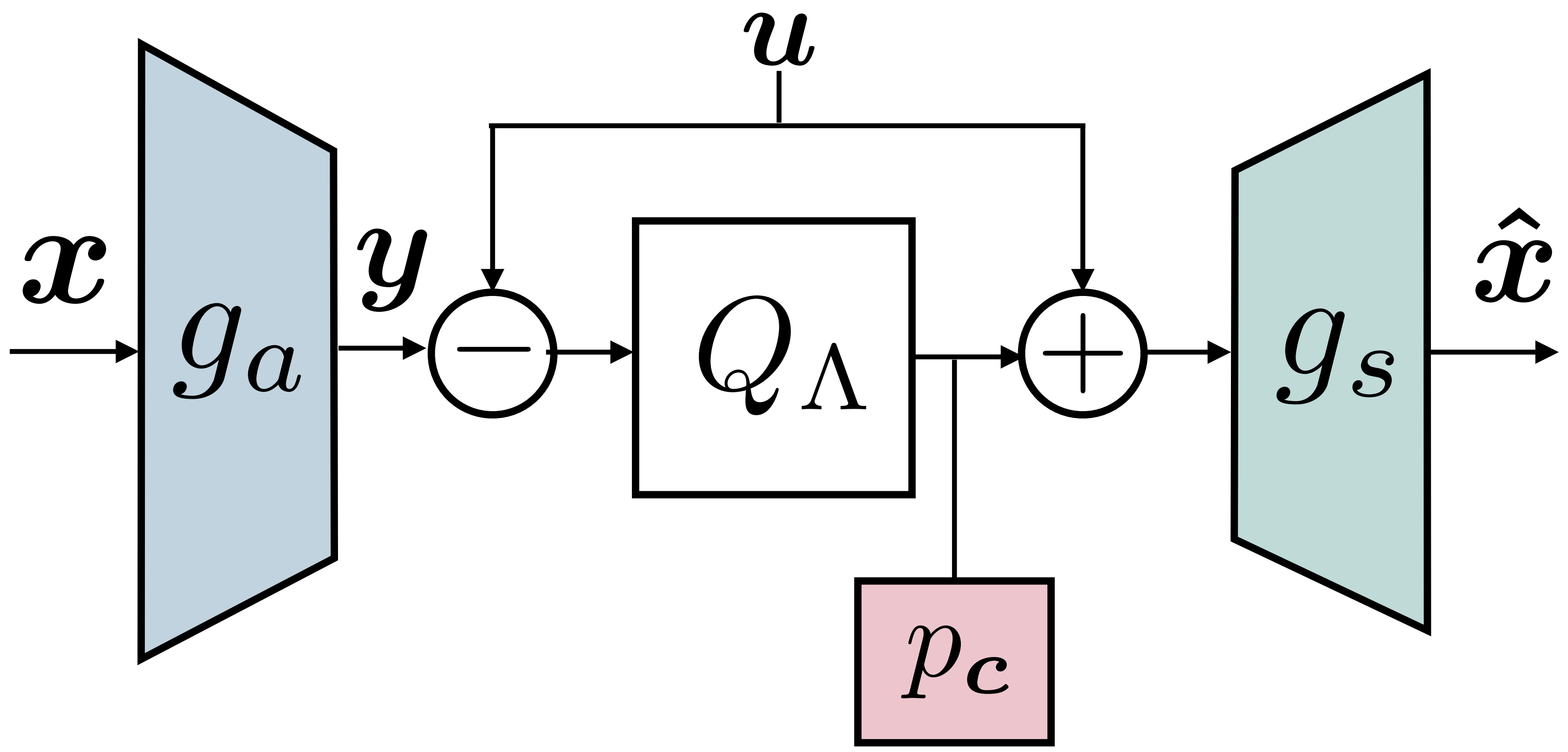
  ✓ Lattice quantization

  ❓ Randomness

LTC with Shared Randomness: Dithering

- Random dither u from the lattice cell, shared between encoder/decoder
- Dithered LQ applied in the latent space:

$$Q_\Lambda(\boldsymbol{y} - \boldsymbol{u}) + \boldsymbol{u}$$



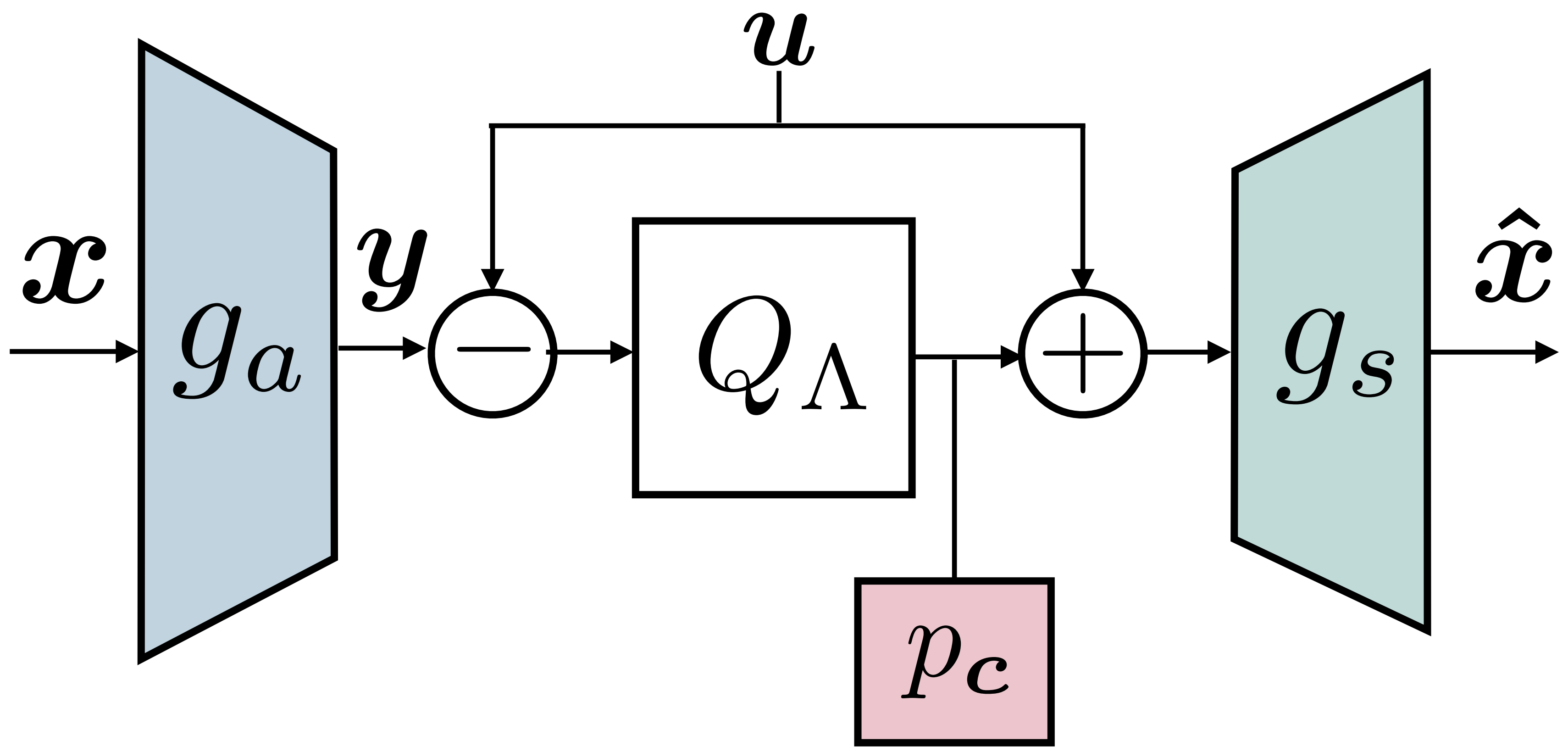
Shared-Dither LTC (SD-LTC)

LTC with Shared Randomness: Dithering

- Random dither u from the lattice cell, shared between encoder/decoder
- Dithered LQ applied in the latent space:

$$Q_\Lambda(\boldsymbol{y} - \boldsymbol{u}) + \boldsymbol{u}$$

- Lattices become sphere-like in high dimensions
- Latent dithered LQ ($Q_\Lambda(\boldsymbol{y} - \boldsymbol{u}) + \boldsymbol{u}$) acts like AWGN channel [Zamir&Feder '96]



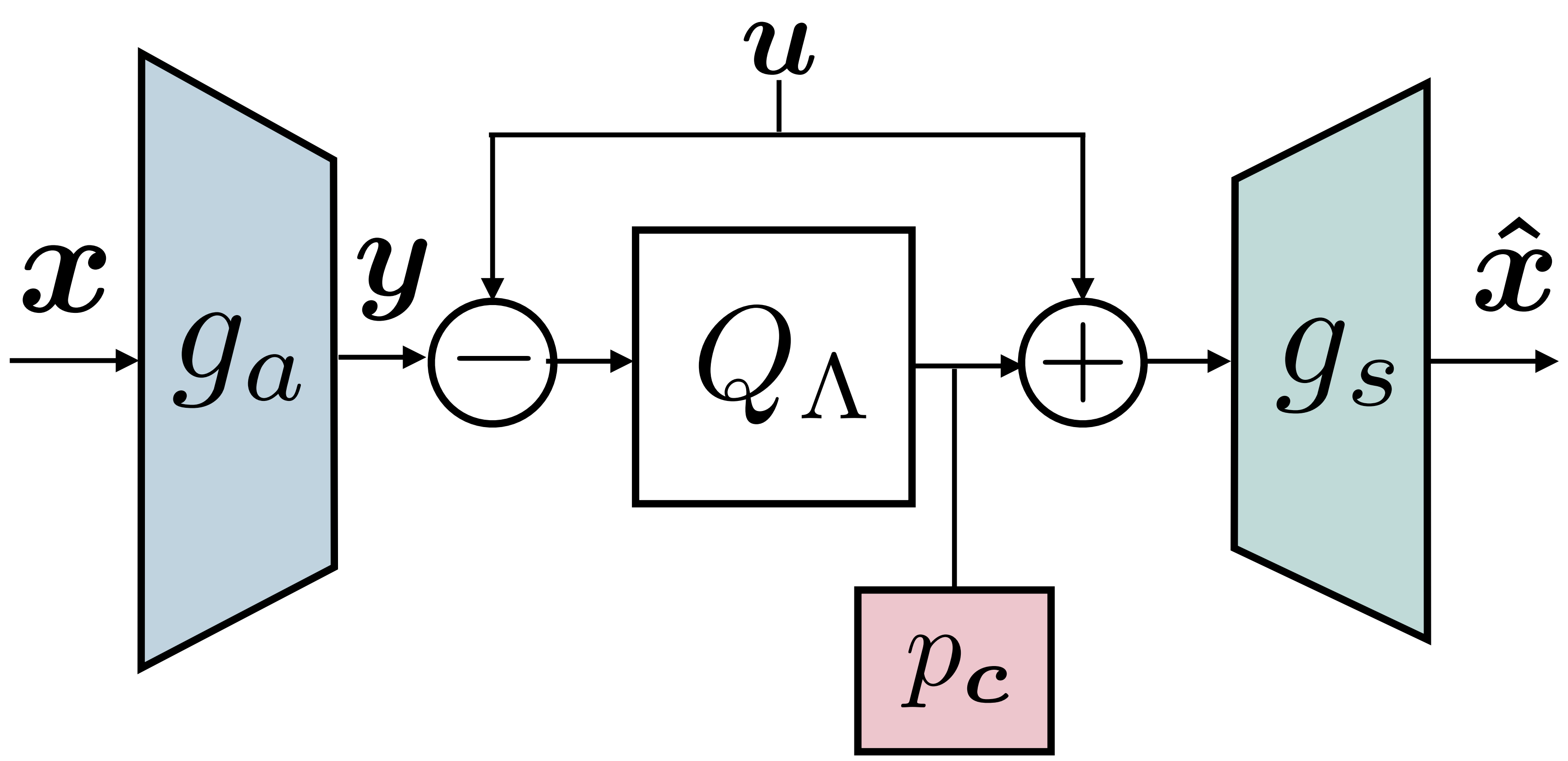
Shared-Dither LTC (SD-LTC)

LTC with Shared Randomness: Dithering

- Random dither u from the lattice cell, shared between encoder/decoder
- Dithered LQ applied in the latent space:

$$Q_\Lambda(\boldsymbol{y} - \boldsymbol{u}) + \boldsymbol{u}$$

- Lattices become sphere-like in high dimensions
- Latent dithered LQ ($Q_\Lambda(\boldsymbol{y} - \boldsymbol{u}) + \boldsymbol{u}$) acts like AWGN channel [Zamir&Feder '96]



Shared-Dither LTC (SD-LTC)

Theorem [Lei,Hassani,SB '25]: Consider an iid Gaussian source, squared error distortion, and a Wasserstein of order 2 for perception measure. SD-LTCs can asymptotically achieve R(D,P).

LTC with No Shared Randomness

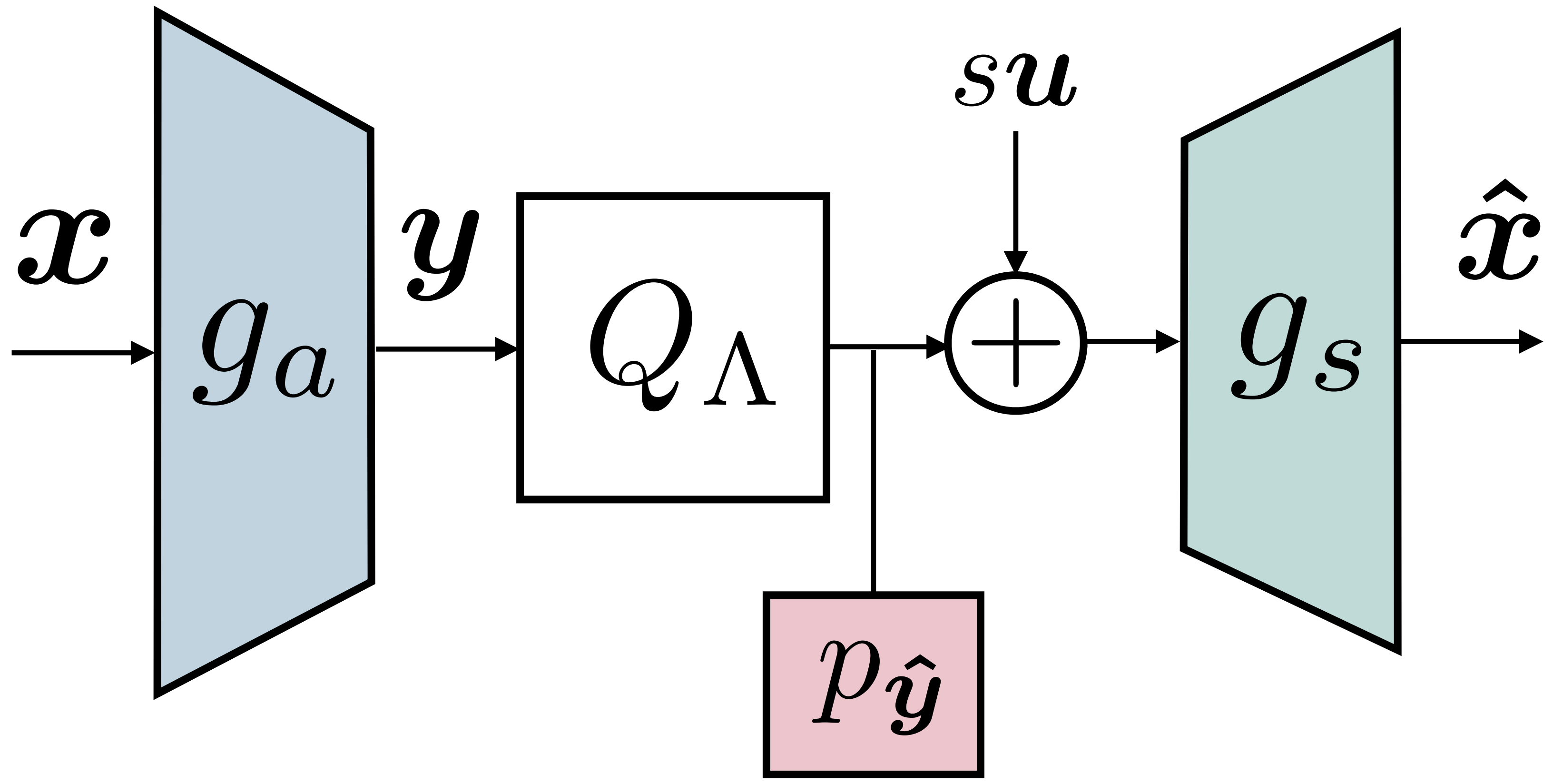
- SD-LTC requires *infinite* shared randomness
    - Not always available
- What if there is no shared randomness

LTC with No Shared Randomness

- SD-LTC requires *infinite* shared randomness
  - Not always available
- What if there is no shared randomness

- Random dither $\boldsymbol{u} \sim \mathrm{Unif}(\mathcal{V}_0)$ **at decoder only**
- Dither applied to quantized latent with scaling:

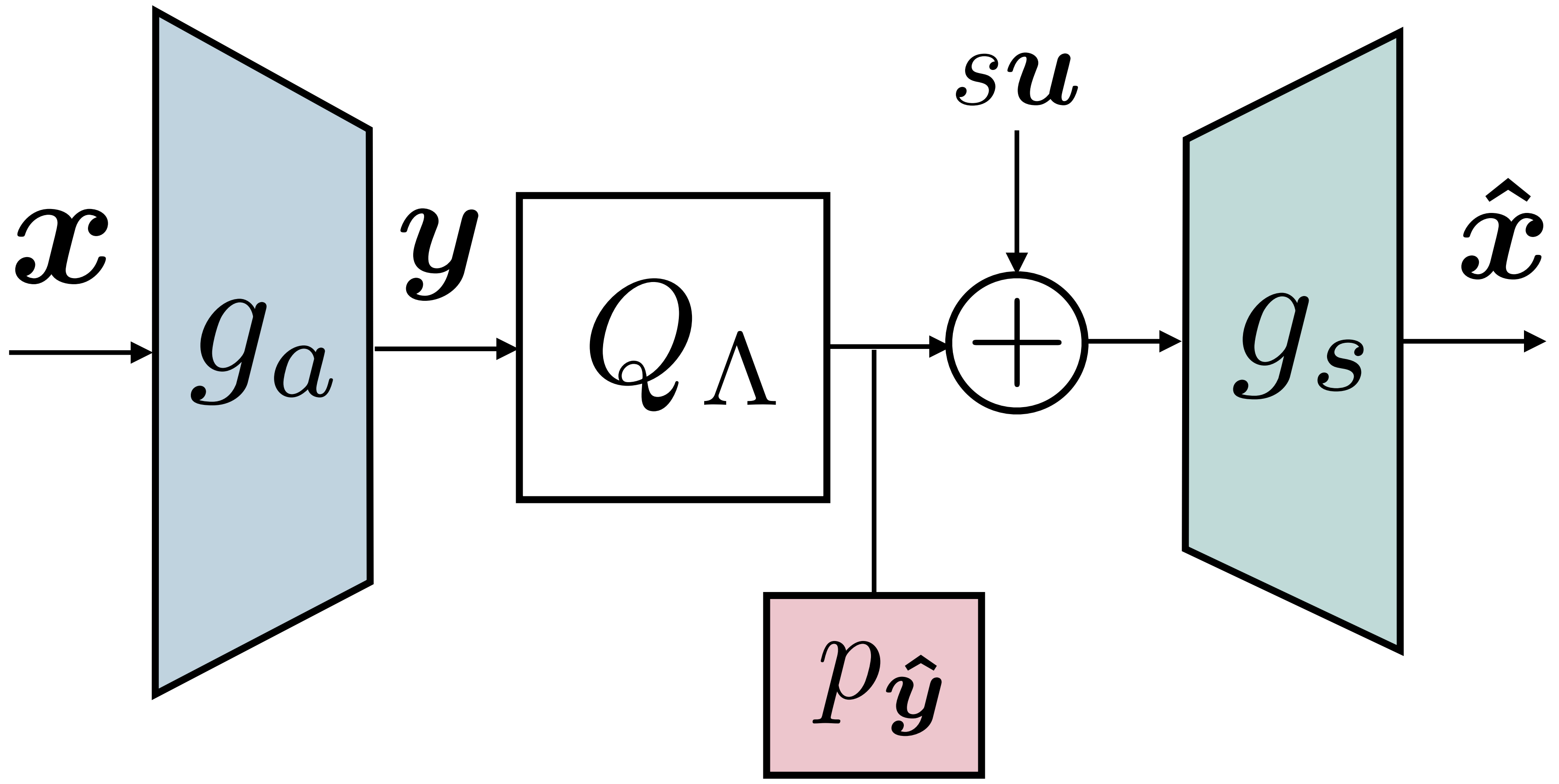
$$Q_\Lambda(\boldsymbol{y}) + s\boldsymbol{u}$$



Private-Dither LTC (PD-LTC)

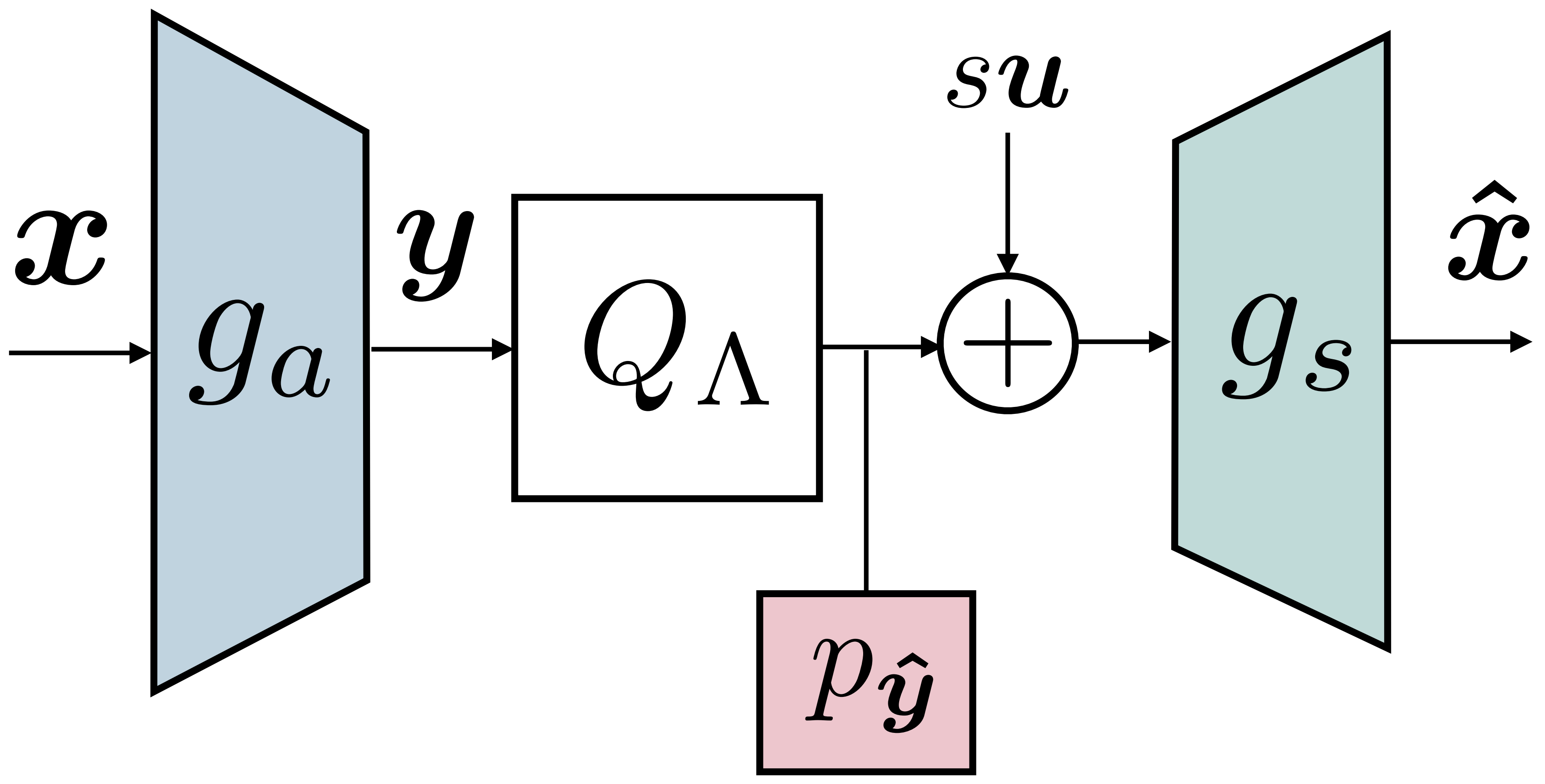
PD-LTC Achievability at $P = 0$

- **Theorem:** PD-LTCs can asymptotically achieve $R(\frac{D}{2}, \infty)$ for iid Gaussians (squared error Wasserstein of order 2 perception).



PD-LTC Achievability at $P = 0$

- **Theorem:** PD-LTCs can asymptotically achieve $R(\frac{D}{2}, \infty)$ for iid Gaussians (squared error Wasserstein of order 2 perception).



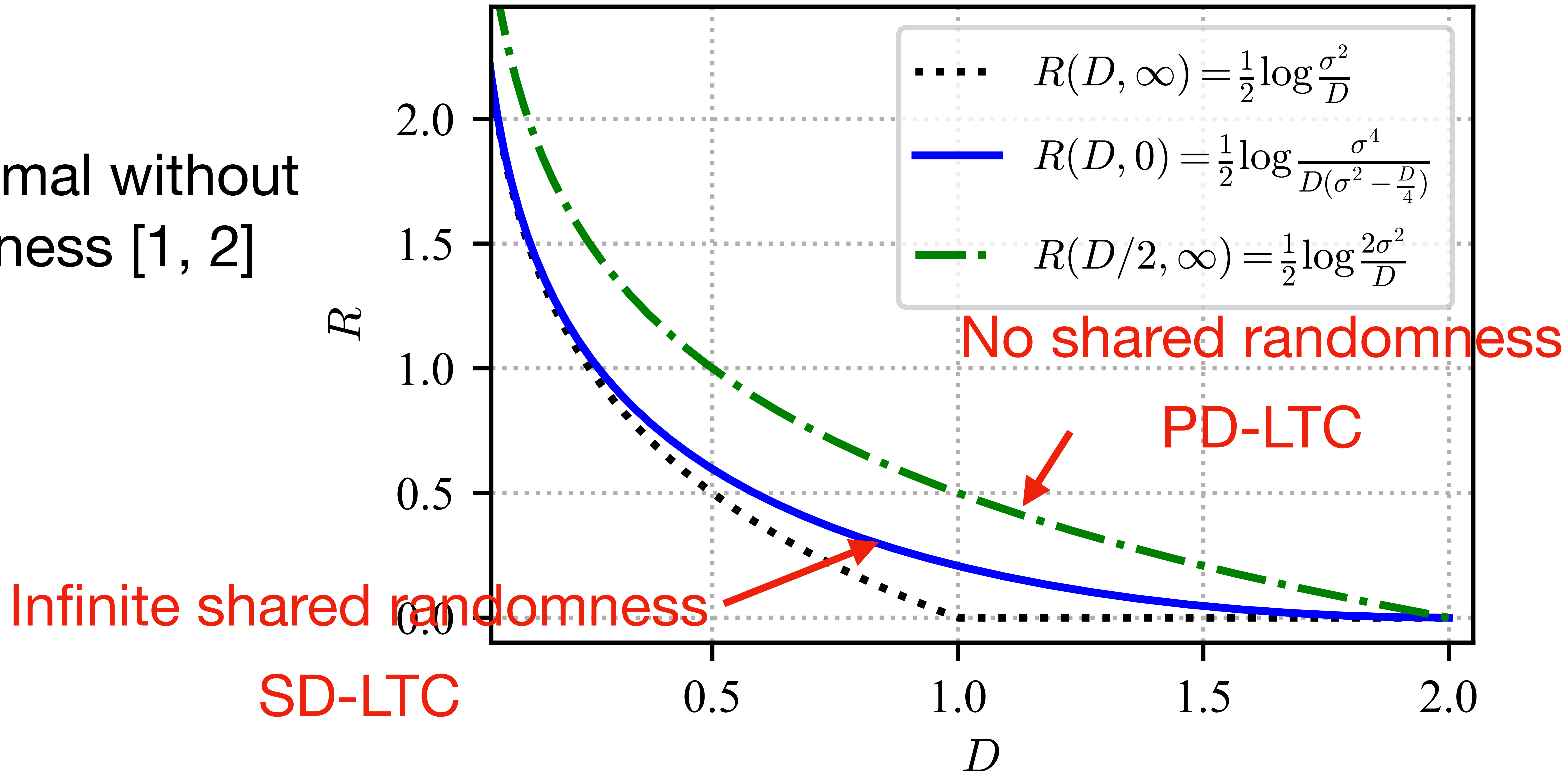
Proof Idea.

- AWGN-equivalence fails
- Proof relies on lattice Gaussian techniques [1] $\quad Q_\Lambda(\boldsymbol{y}) \approx \text{Lattice Gaussian}$
- $s = \frac{\sigma}{\sqrt{\sigma^2 - D/2}} \implies$ enforces perception constraint

[1] C. Ling and J.-C. Belfiore. Achieving awgn channel capacity with lattice gaussian coding. IEEE Trans. Inf. Theory, 2014.

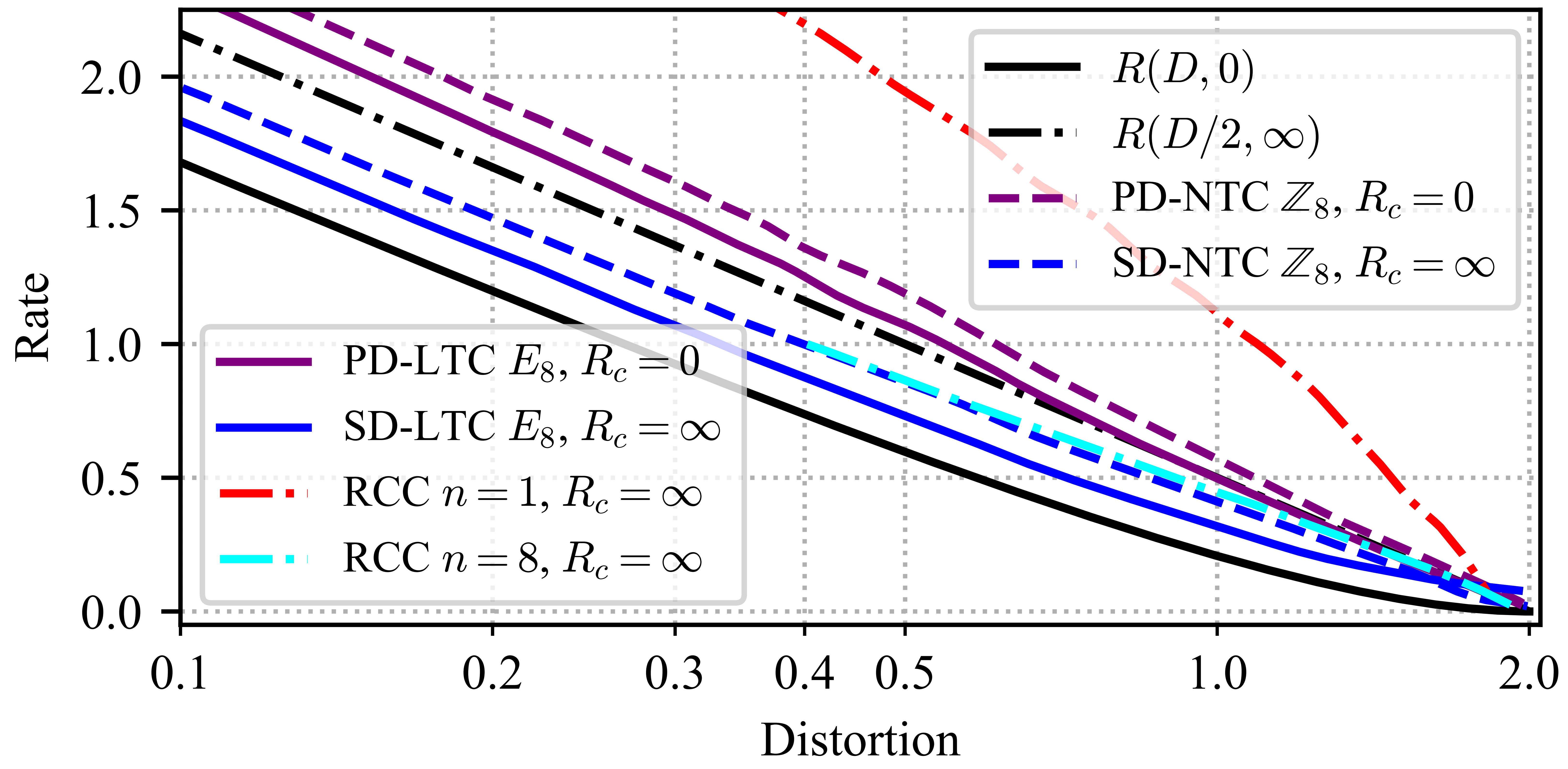
Comparing Fundamental Limits

- Consider $P = 0$
- $R(D/2, \infty)$ optimal without shared randomness [1, 2]



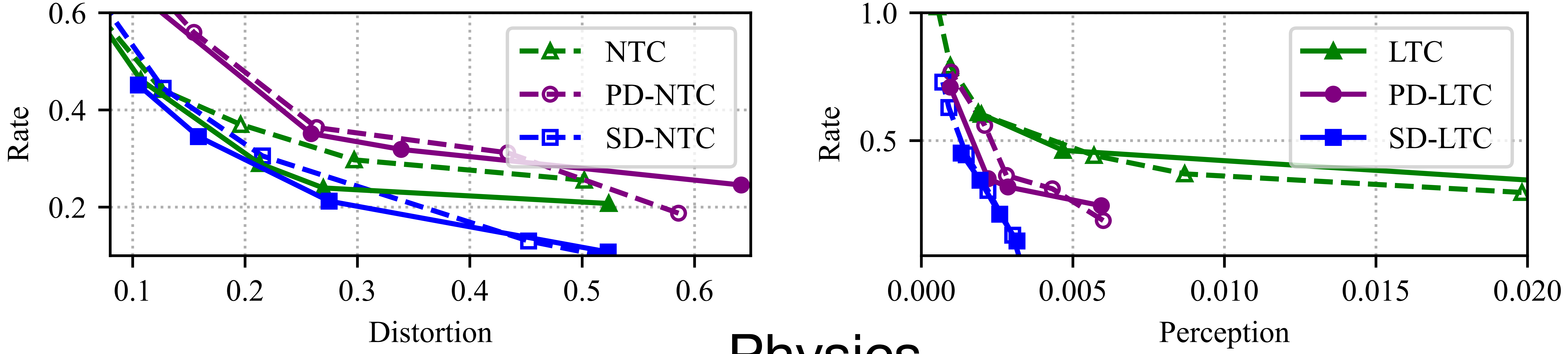
[1] N. Saldi, T. Linder, and S. Yüksel. Output constrained lossy source coding with limited common randomness. IEEE Trans. Inf. Theory 2015.
[2] A. B Wagner. The rate-distortion-perception tradeoff: The role of common randomness. arXiv 2022.

Experimental Results: Gaussian

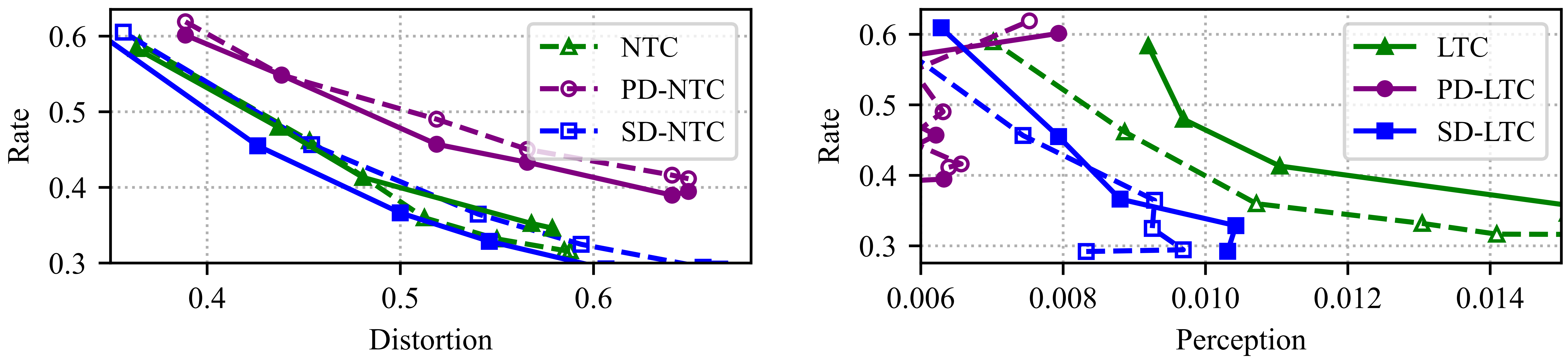


Experimental Results: Real-World Sources

- Speech and Physics sources [Yang & Mandt, 2022]



Physics



Speech

Conclusion & Future Work

- We proposed neural compressors that provide VQ-type solutions, allow shared randomness into the design, have low complexity, and performance guarantees for Gaussian sources.

- Generalizing the analysis of PD-LTC to P>0
- Generalizing the solution to limited randomness
- LTC for distributed compression, in line with [Ozyilkan et al ’23]